Symposium on Statistical Bases for Public Health Decision Making: From Exploration to Modelling

Guest Editors: Kenneth H. Falter, Donald R. Betts, Deborah B. Rolka, Henry R. Rolka, W. Karl Sieber

Statistics in Medicine

Volume 18 Number 23 15 December 1999

D. Machin
R. B. D'Agostino

SMEDDA 18(23) 3159–3376 (1999)
ISSN 0277-6715

STATISTICS IN MEDICINE

CONTENTS

VOLUME 18, ISSUE No. 23 15 December 1999

Symposium on Statistical Bases for Public Health Decision Making: From Exploration to Modelling
Guest Editors: Kenneth H. Falter, Donald R. Betts, Deborah B. Rolka, Henry R. Rolka, W. Karl Sieber

Preface 3159

Staff and Committees 3161

Introduction and Welcome 3165

Simultaneous Smoothing and Adjusting Mortality Rates in U.S. Counties: Melanoma in White Females and White Males 3167
K. Kafadar

Model-based Small Area Estimates of Overweight Prevalence Using Sample Selection Adjustment 3189
D. Malec, W. W. Davis and X. Cao

Application of a Weighted Head-banging Algorithm to Mortality Data Maps 3201
M. Mungiole, L. W. Pickle and K. Hansen Simonson

Exploring Spatial Patterns of Mortality: The New *Atlas of United States Mortality* 3211
L. W. Pickle, M. Mungiole, G. K. Jones and A. A. White

All Maps of Parameter Estimates are Misleading 3221
A. Gelman and P. N. Price

Modelling for Cost-effectiveness Analysis 3235
L. B. Russell

Constructing Confidence Intervals for Cost-effectiveness Ratios: An Evaluation of Parametric and Non-parametric Techniques Using Monte Carlo Simulation 3245
A. H. Briggs, C. Z. Mooney and D. E. Wonderling

Evaluating the Cost-effectiveness of Vaccination Programmes: A Dynamic Perspective 3263
W. J. Edmunds, G. F. Medley and D. J. Nokes

A Monitoring System for Detecting Aberrations in Public Health Surveillance Reports 3283
G. D. Williamson and G. Weatherby Hudson

The Cumulative q Interval Curve as a Starting Point in Disease Cluster Investigation 3299
R. Chen

A Study of the Average Run Length Characteristics of the National Notifiable Diseases Surveillance System 3309
L. VanBrackle and G. D. Williamson

(*continued on inside back cover*)

0277-6715(19991215)18:23;1-U

(*continued from outside back cover*)

Estimating Genetic Influence on Disease from Population-based Case-control Data: Application to Cancers of the Breast and Ovary 3321
G. Gong and A. S. Whittemore

An Application of Lifetime Models in Estimation of Expected Length of Stay of Patients in Hospital with Complexity and Age Adjustment 3337
J. Li

Multi-criteria Decision Making – An Approach to Setting Priorities in Health Care 3345
F. Fonseca Nobre, L. T. Ferreira Trotta and L. F. A. Monteiro Gomes

Recurrent Injury Event–Time Analysis 3355
J. T. Wassell, W. C. Wojciechowski and D. D. Landen

Presentations 3365

Closing Remarks 3373

STATISTICS IN MEDICINE
Statist. Med. **18**, 3159–3160 (1999)

PREFACE[†]

The symposium 'Statistical bases for public health decision making: from exploration to modeling' is the sixth in a series of statistical methods symposia sponsored by the Centers for Disease Control and Prevention (CDC) and the Agency for Toxic Substances and Disease Registry (ATSDR) since 1988. Previous symposia were 'Small area statistics in public health', 'Quantitative methods for utilization of multi-source data in public health', 'Statistical methods for evaluation of intervention and prevention strategies', 'Clustering and health events', and 'Statistics in surveillance'. These symposia underscore the importance of statistical and other analytic methods in understanding and preventing disease, injury and other unhealthy conditions.

With government downsizing and increasing competition for funds, it has become more important than ever to base public health decisions on solid statistical and quantitative evidence. Proper data analysis and presentation (including modelling, risk-benefit analysis, smoothing techniques and easily understandable graphic presentation of results), as well as properly designed and executed studies and sample surveys, are important facets of providing such statistical bases, which is why these topics were selected as symposium sub-themes.

The sixth symposium consisted of a mixture of plenary sessions, parallel sessions, poster sessions, and a short course. Almost 300 participants from six countries and the U.S.A., representing academia, local, state and federal governments, and other health-related organizations, attended and contributed to all these sessions.

Drs. Sallie Keller-McNulty and Karen Kafadar very capably taught an excellent course on 'Exploratory data analysis and graphical methods for public health decision making' to more than 90 attendees on the day prior to the main symposium.

The symposium's opening plenary session was on statistical bases for public health decision making. Dr. David Sencer, former CDC Director, spoke on the question 'Are statistics necessary for public health decisions?' He critiqued the statement 'public policy decisions are often criticized for relying on political motives rather than on sound data' and gave examples to illustrate both the truth and fallacy of the statement. He was faced with a potential influenza epidemic and a paucity of available data when he had to decide whether or not to authorize use of swine flu vaccine. This was a public health decision with broad implications that had to be made under great uncertainty. He stated that 'at times, epidemiological intuition coupled with incomplete data will lead to decisions that protect the public. Policy makers should not only be aware of the benefits of statistical reasoning and modeling, but should also have an appreciation of history'.

Dr. Edward Sondik, director of CDC's National Center for Health Statistics and formerly of the National Cancer Institute, National Institutes of Health, gave a presentation titled 'Use of cancer data to influence public health policy'. He discussed the meta-analyses performed to determine the effectiveness of breast cancer screening for women in their forties in deciding

Published in 1999 by John Wiley & Sons, Ltd.

whether or not to recommend an annual mammogram for women under age 50. He reviewed the studies done with an eye on understanding how conflicting opinions can draw on the same statistical data for their conclusions. The evidence was at the borderline of statistical significance, depending on which studies were included in or excluded from the meta-analysis. Owing to continuing uncertainty, no public health decision has been made.

Dr. J. Lee Annest of CDC's National Center for Injury Prevention and Control gave a talk titled 'How national data on blood lead levels served as the scientific basis for federal policy decisions to limit the use of lead in gasoline'. He described how national survey data on blood lead levels were analysed and used to influence federal regulatory actions to restrict the use of lead in gasoline. He investigated trends in average blood lead levels for persons aged 6 months to 74 years in relation to total lead used in gasoline production. The drop in average blood lead levels by 6-month periods was highly correlated with the reduction in total lead used in gasoline per 6 months. A simple presentation of these data backed by a detailed statistical evaluation had a major impact on the decision by federal regulators to restrict the use of lead as an octane booster in gasoline.

Two additional plenary sessions, 'EDA (exploratory data analysis) and survey data', and 'Cost-effectiveness and modelling', served to introduce the sub-themes of the symposium. Eight parallel sessions and the poster sessions further elucidated the many aspects of the symposium's themes.

This symposium was enriched by the efforts of its co-sponsors: the Atlanta Chapter of the American Statistical Association; the Biostatistics Department, Rollins School of Public Health; the Department of Psychiatry and Behavioral Sciences, School of Medicine, Emory University; and the Biostatistics Department, University of Cincinnati. We thank them for their significant contributions to the scientific programme and administrative aspects of the symposium. We thank all speakers, moderators and poster presenters for their excellent scientific contributions to the symposium; Dr. Theodore Colton of Boston University, one of the founding editors of *Statistics in Medicine*, for his continuing support of these symposia; and Dr. Donna Stroup, CDC Epidemiology Program Office, who capably filled in for Dr. Colton in presenting an excellent summary of the symposium and the implications of what was presented.

My special thanks go to the symposium Planning Committee chairs and members and to the editorial board and reviewers of the symposium abstracts and of papers submitted for this issue. Without their many hours of effort, this symposium would not have been the success that it was! Thanks also to our meeting planner, Bachmann and Associates, Inc., for handling the details of registration, publicity, site selection, travel and logistic challenges in a manner that kept the meeting flowing smoothly.

In this issue, we present a selection of outstanding papers that convey the flavour and quality of the sixth CDC and ATSDR symposium on statistical methods. We hope you enjoy reading them as much as we have.

K. H. FALTER
Centers for Disease Control and Prevention

STATISTICS IN MEDICINE
Statist. Med. **18**, 3165 (1999)

INTRODUCTION AND WELCOME[†]

I am delighted to have the opportunity to welcome you all to the symposium. During my career at CDC as an epidemiologist, I had the pleasure of working closely with many statisticians, and when I served as Associate Director for Science, the Statistical Advisory Group educated me about the importance of statisticians, and statistical continuing education for CDC.

This is the sixth biannual statistics symposium sponsored by CDC/ATSDR and other groups; as has occurred previously, we will benefit from the state-of-the-art plenary lectures and invited papers and poster presentations, and always when professionals get together there is the excitement of exchanges of ideas and cross-fertilization across fields and disciplines. The topic of this symposium is particularly important, as it reflects an increasing recognition of how statistics can contribute to the whole range of our public health activities. Certainly you will be discussing traditional statistical roles in data analysis, exploratory data analysis, modelling and survey-related issues. Two session topics represent a broader perspective as to how statisticians might apply their methods for public health benefit. The first session on 'Influence of statistical methods on the development and implementation of public health policy' directly acknowledges that statistics can contribute not just development of data but use of those data and analyses for policy development.

I presume that many of the papers in this session will focus on prevention effectiveness and some of the more traditional uses of decision analysis and cost effectiveness analysis, but I encourage you to push the envelope and think creatively about what methods might apply to broader programme evaluations, as well as evaluations of specific interventions. Programme evaluation has always been of importance to CDC, but it has not always been done as rigorously as it might be. It is difficult where individual randomization is usually not feasible. That presents a challenge to think about how can we approach programme evaluation in a rigorous way. Please keep that challenge in mind as you look toward the future when you can apply what you learn at this symposium.

We appreciate the broad range of expertise and experience in the audience, including participation from government agencies, academia and private industry. We are also pleased at the international representation, with presenters from the United States, Brazil, Canada, England, Germany, Israel and Japan.

In closing, I would like to emphasize that this symposium, as well as the previous symposia, and the biannual symposia in the future, are really important for us at CDC in carrying out our mission. Statisticians are integrally involved in the practice of public health science, and it is crucial at CDC that our statisticians use state-of-the-art methodology and have the opportunity for cross-fertilization with developments in academia and industrial statistical practice. Our statisticians need to be particularly creative in applying methods developed in different areas to public health issues or developing new approaches. I trust that this symposium will evoke new ideas for how statisticians can be involved in helping to improve public health policy and practice.

CLAIRE V. BROOME
Centers for Disease Control and Prevention and
Agency for Toxic Substances and Disease Registry

Published in 1999 by John Wiley & Sons, Ltd.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3167–3188 (1999)

SIMULTANEOUS SMOOTHING AND ADJUSTING MORTALITY RATES IN U.S. COUNTIES: MELANOMA IN WHITE FEMALES AND WHITE MALES

KAREN KAFADAR*

Department of Mathematics, Box 170, University of Colorado-Denver, Denver, Colorado 80217-3364, U.S.A.

SUMMARY

Detecting patterns in health-related data for geographic areas is facilitated with the use of exploratory methods, especially smoothing. In addition, these data often must be adjusted for known prognostic factors such as age and gender. The analysis in this paper focuses on mortality rates due to malignant melanoma in White males and White females; these data are adjusted for both age and latitude, separately for males and females, and then smoothed using (a) a non-linear smoother known as weighted head-banging, and (b) a new method that incorporates the adjustment and the smoothing simultaneously. Maps of the continental United States show regions of high rates, even after having adjusted for age and latitude, and suggest the possibility of other variables that may influence the rates. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

A major challenge in interpreting public health data is the development of appropriate summary and display techniques that can provide insight into possible causes, effects and trends in the vast amounts of data. Exploratory methods can be valuable tools in this process, primarily in the initial stages of both model formulation and verification of assumptions often required by formal parametric methods. The purpose of these methods is to provide insight as opposed to precise estimates of location, spread or trends.

Smoothing is an exploratory method that is particularly valuable for geographically-defined data, since patterns are difficult to decipher in two dimensions, especially when the data are subjected to noise from various sources (inaccurate records or counts of cases and populations, mislocated cases, etc.). Various papers have proposed empirical Bayes methods for incidence and mortality rates, often in the context of estimating the true rates by reducing noise and achieving greater homogeneity among their variances; some authors subsequently have recognized that the stabilized rates can be mapped and that the subsequent map is 'smoother' than the map of the raw rates. This approach is apparent in the papers by Manton *et al.*[1] for analysing U.S. cancer mortality, Marshall[2] in the analysis of cancer mortality in England and Wales, Clayton and

* Correspondence to: Karen Kafadar, Department of Mathematics, Box 170, University of Colorado-Denver, Denver, Colorado 80217-3364, U.S.A. E-mail: kk@math.cudenver.edu

Contract/grant sponsor: National Science Foundation
Contract/grant number: DMS 95-10435

CCC 0277–6715/99/233167–22$17.50

Kaldor[3] and Kaldor and Clayton[4] for lip cancer mortality in Scotland, Cislaghi *et al.*[5] for breast and lung cancer mortality in Italy, Besag *et al.*[6] for multiple myeloma mortality in France, Cressie[7] for sudden infant death syndrome in North Carolina, and Devine *et al.*[8–10] for lung and brain cancer mortality in Ohio and burn-related mortality in the U.S. A few papers have focused on the smoothing technology directly and have gained insights from its application: Bloomfield,[11] Carr and Pickle,[12] Cressie and Read,[13] Kafadar,[14,15] and Mungiole *et al.*[16] Two-dimensional smoothers have been applied in many other contexts besides public health such as economics,[17] geology,[18] and atmospheric science.[19]

The purpose of this paper is to describe two general classes of smoothers and illustrate the application of specific members of each class on melanoma mortality data. Smoothers can be very useful in analysing geographical data in public health, particularly when those data are in the form of rates or adjusted for obvious prognostic factors such as age, environmental exposures and demographic variables. Tukey[20] wrote of 'ecological exploration' oriented towards revealing patterns which can hide amidst noisy data and, more usefully, suggest important variables for adjustment. Most researchers recognize the need to standardize mortality or incidence rates before mapping, either for age or as a standardized mortality/incidence ratio (SMR/SIR). However, just as an examination of the residuals from a straight line fit can highlight hidden patterns or variables of interest, the adjustment of rates for strategic variables allows one to view them as if they were under a microscope. Twenty years ago, Tukey[21] advised:

> Where we are concerned with events occurring to people, few if any of us would plot raw event counts or raw event rates. Instead, almost all of us would plot rates adjusted for age and sex … What we are accustomed to take for granted as far as age and sex goes, we should take equally for granted as far as such variables as
>
> (i) size of places
> (ii) general/special ethnic distribution
> (iii) industry/office/agriculture split
> (iv) distribution of types of industry
> (v) income profile of family units
> (vi) you name it.

Interestingly, authors often continue to map raw rates, even when risk factors are acknowledged (for example, reference 5, p. 2367). In situations where risk factors have not been confirmed or even identified, the results of smoothing the age-adjusted rates can suggest variables that might be worthwhile to use as adjusters. Once the data have been so adjusted, smoothing can again highlight further patterns and subsequent adjusters. This iterative approach – smooth, plot, adjust; resmooth, replot, readjust – is much within the spirit of exploratory data analysis[22] and provides an analysis that can be interpreted at each stage. Moreover, different smoothers may be needed for different purposes; linear smoothers expose extremely broad, non-specific trends, and non-linear smoothers identify sharp distinctions between regions.

Section 2 describes general goals and purposes of smoothing, and Section 3 provides a brief introduction to two classes of smoothers that may be needed to achieve them. Discussion centres around three specific smoothers, two linear and one non-linear, because they are applied to malignant melanoma mortality rates in U.S. counties, described in Section 4. Section 5 presents results of smoothing the age-adjusted rates, followed by methods for further adjustment for latitude and simultaneous smoothing of the residuals in Section 6. Section 7 describes the results

of applying the methods, and Section 8 concludes with a summary and recommendations for choosing different smoothers to achieve different objectives.

2. OBJECTIVES OF SMOOTHING AND GOALS OF SMOOTHERS

Tukey[22] advocated smoothing as a tool in the exploratory analysis of data and illustrated its use on data sequenced by one variable such as time; to this day, most smoothing methodologies are described in the context of one-dimensional data. The reasons for smoothing one-dimensional data apply even more forcefully to the case for two- and higher-dimensional data, since patterns therein are typically more difficult to decipher. Since measurement error and unusual values ('outliers') can mask features in the data, such as trends, peaks, troughs, or ridges, smoothing can help to reduce the distracting noise and enable one to see the patterns. Some of the objectives for smoothing techniques are:

(i) to enhance the relationship between location and outcome;
(ii) to reduce the variance in measurement errors, thereby allowing the underlying trend to be seen more easily;
(iii) to reduce attention to unusual values or outliers;
(iv) to examine patterns in the residuals that can be revealed once the smoothed trend has been removed;
(v) to minimize the effect of aggregation in what may be a summary data point across an entire region (census tract, county, district, state economic area).

These objectives differ from those for prediction and interpolation, where often the goal is to derive a value at some intermediate location other than where data are taken, or for model estimation, since we plan to let the data propose the model rather than impose a model on the data. Because errors affect all data to some extent, and since the eye almost always needs assistance in deciphering patterns, some form of smoothing is almost always advantageous. Smoothed values have less variance that the original data, but they also have more bias, so the trade-off between bias and variance must be considered when choosing a smoother. For the most part, however, the inclusion of some bias is often well worth the reduction of variance if the result is a better understanding and a faithful rendition of the underlying trends in the data.

3. TYPES OF SMOOTHERS

Smoothing algorithms can be classified broadly as *linear*, such as moving (weighted) averages,[23] local polynomial regression,[17] spatial splines,[24] or, for data specifically in the form of rates, empirical Bayes smoothers,[3] and *non-linear*, such as Tukey's median filter,[25] median polish smoother[13,22] (also called separable median filter in the engineering literature[26]), and head-banging.[18,27] To describe these classes, suppose there are n observations, $\{y_1, \ldots, y_n\}$ observed at locations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Occasionally a particular case may be associated with exact geographical co-ordinates, but more commonly the data consist of only counts of cases and of persons or person-years in a given unit over a given time interval, in which case $\mathbf{x}_i$ represents some convenient centroid for the ith area. The observation y_i is modelled as a sum of the true function, say $F(\mathbf{x}_i)$, and a noise term, say ε_i, which may represent a minor departure from $F(\mathbf{x}_i)$ such as

measurement error or a substantial error because of the unusual nature of the ith area. Some performance characteristics of a good smoother are:

(a) it should recreate F as accurately as possible;
(b) it should recapture linear surfaces that have no error;
(c) its performance should not be substantially impaired if the data are not evenly spaced;
(d) its output should be 'smooth' where F is, without attempting to smooth over obvious breaks;
(e) unusual values, unsupported by neighbours, should stand out clearly in the residuals, not in the smooth.

Linear smoothers can always be expressed as a linear function of the observed values; if $\tilde{y}_i$ is the smoothed value of y_i at location $\mathbf{x}_i$, then $\tilde{y}_i$ can be expressed as

$$\tilde{y}_i = \sum_{j=1}^{n} a_{ij} y_j \Big/ \sum_{j=1}^{n} a_{ij}.$$

The weights of $\{y_1, \ldots, y_n\}$ typically will depend upon the target point being smoothed. For example, a simple disk average smoother has $a_{ij} = 0$ if the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ exceeds a certain threshold (radius of the disk) and 1 otherwise. *Kernel smoothing*[28] arises when the weights a_{ij} are a function, say $K(\cdot)$, of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$; for example, $K(\|\mathbf{x}_i - \mathbf{x}_j\|/h)$, where h is a *bandwidth* (larger $h \rightarrow$ smoother result). Spline-based smoothers[24,29] are designed to fit the data well but also be smooth, and thus to balance explicitly between characteristics (a) and (d) above. Asymptotically, smoothing splines can be expressed as linear functions of the data, and hence they fall into the class of linear smoothers (reference 28, Section 5.3, p. 171).

A popular smoother is *loess*, a local polynomial regression smoother.[17,30] The smoothed value is the prediction from a regression surface over the point $(\mathbf{x}_i, y_i)$ and its neighbours, where the neighbours constitute a fixed fraction, f, of the n sample points that are closest to $\mathbf{x}_i$ and the weights involve a function of the distance of these neighbours to $\mathbf{x}_i$. Simonoff[28] shows that the local polynomial regression smoothers are likewise linear smoothers (Section 5.2.1, pp. 138–139). The increased complexity of the method allows greater flexibility in the smoothed output than simple moving (weighted) averages. The weights for the former actually involve the co-ordinates $\mathbf{x}_i$, while those of the latter usually depend only on interpoint distances. These weights can also involve the variances of the points, for example, person-years, so that points that are further away but are very stable can have more influence than y_i itself on $\tilde{y}_i$. An appropriate choice for the fraction f can thus capture features in the surface such as inclines, hills, or valleys in any direction – but the appropriate choice for f is not a trivial problem. Simonoff[28] (Chapter 5) discusses bandwidth selection rules for various linear smoothers, including spline smoothers, kernel smoothers and loess.

Kafadar[15] suggested a linear smoother appropriate specifically for data in the form of rates. As its application to data in Section 5 will illustrate, this age-specific rates smoother is very useful for suggesting possible adjusters, primarily because it has a tendency to oversmooth. Let $r_{km} = D_{km}/P_{km}$ denote the age-specific rate in county k, $(k = 1, \ldots, n)$ and age group m $(m = 1, \ldots, 18)$, where D_{km} and P_{km} are the numbers of deaths and person years, respectively, and let d_{jk} denote the (Euclidean or great circle) distance between counties j and k. The smoothed rate is

$$\tilde{r}_k = \sum_{m=1}^{18} = \pi_m D^*_{km}/P^*_{km}$$

where

$$D^*_{km} = \sum_{j=1}^{N} w_{jk} D_{jm}$$

$$P^*_{km} = \sum_{j=1}^{N} w_{jk} P_{jm}$$

π_m = proportion of the standard population in age group m

$$w_{jk} = [1 - (d_{jk}/d_{\max})^2\}^2 \quad d_{jk} \leqslant d_{\max}$$
$$= 0 \qquad\qquad\qquad\qquad d_{jk} > d_{\max}.$$

The parameter $d_{\max}$ controls the rate of smoothing; larger values of $d_{\max}$ provide smoother results.

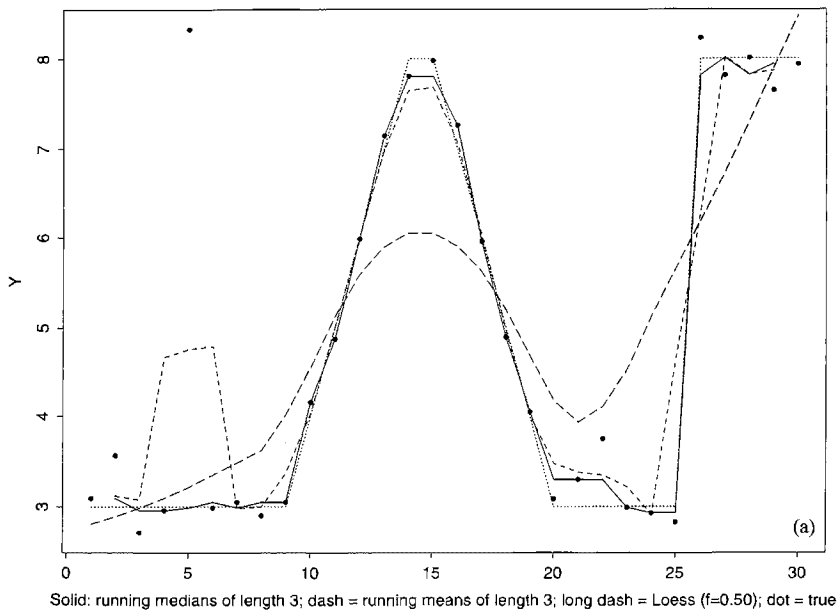
Various weight functions other than that suggested above (called the bisquare weighting function[31]) can be used, but in practice the particular choice is not terribly crucial, so long as it is sensible. In the application that follows, the weights are a function not only of distance, but also of the standard error of the rates:

$$w_{jk} = \left(\sum_{m=1}^{18} P_{jm} \right)^{1/2} [1 - (d_{jk}/d_{\max})^2]^2 \quad d_{jk} \leqslant d_{\max}$$
$$= 0 \qquad\qquad\qquad\qquad\qquad\qquad d_{jk} > d_{\max} \tag{1}$$

normalized so that the sum of the weights is one. Note that, with this smoother, deaths and populations are 'smoothed' separately, and then the resulting smoothed age-specific rates are adjusted for age accordingly. Theoretical performance of this smoother is detailed elsewhere.[15]

If the underlying surface is actually non-linear, the linearity of a smoother need not impede its ability to provide a decent approximation to it, and linearity does have the advantages of straightforward formulation and implementation. However, linear smoothers tend to be somewhat unsatisfying in situations with abrupt features, extreme values, or outliers. For example, a sudden shift in the data due to a mountain range or river may be geographically meaningful in explaining a change in rates on either side, but a linear smoother will tend to smooth over this feature and make it appear as a gradual change. While smoothers are indeed supposed to 'smooth', they should also be faithful to the data, and the suggestion of a gradual change in such a situation would not be valid. The same principle applies to the phenomenon of peaks and troughs; linear smoothers will squash peaks and raise troughs, and, in some instances, obliterate these features altogether. Neither characteristic is in line with (a), recreating F as accurately as possible. Another problem is their unfortunate response to unusual values; while a high rate for, say, New York City, may be correct, it should not affect our impression of the overall trend, unless neighbours support its value with equally high rates. However, instead of ignoring such values, linear smoothers will give them weight, a phenomenon that Tukey calls 'tenting'.[22] Weights based on variances of the points can help, but an outlier in a place the size of New York City is likely to exert large influence on other smoothed values. Thus, while linear smoothers have some of the desirable features of a smoother (for example, (b), (d)), they are far from ideal.

For abrupt features, median-based (hence non-linear) smoothers provide more faithful renditions of the underlying patterns. Because they use medians, abrupt features will be retained far better than with linear smoothers. Moreover, an unusual value that is smeared across several values by a linear smoother will be ignored by a median-based smoother, and its unusual value will show up clearly in the residuals. Figure 1(*a*) illustrates this phenomenon using a simulated



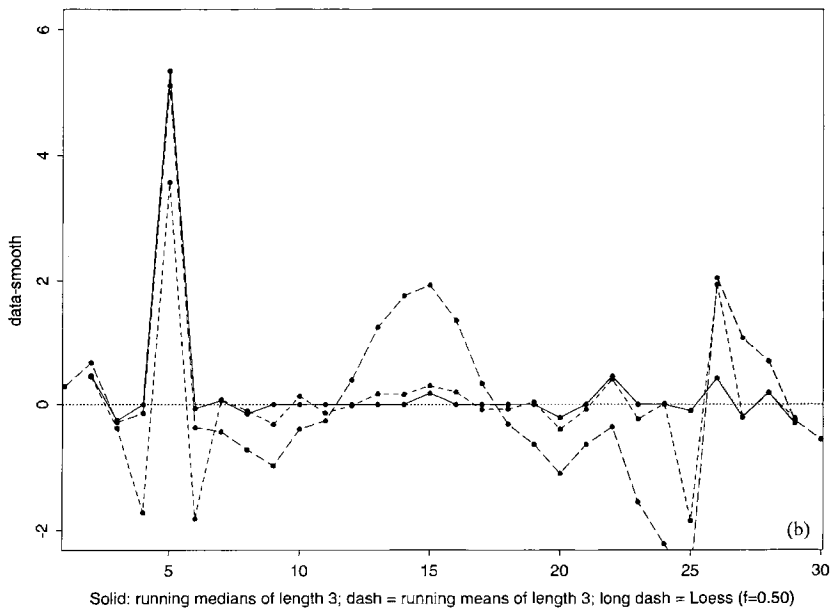


Figure 1. Simulated time series illustrating linear and non-linear smoothers in one dimension. (*a*) The data points are obtained from the true function, shown as a dotted line, with error: solid line, running medians of length 3 (non-linear smoother); short dashed line, running means of length 3 (linear smoother); long dashed line, linear loess smoother, span $f = 0{\cdot}50$. (*b*) Residuals = data in panel (*a*) minus the smoothed value: solid line, running medians of length 3; short dashed line, running means of length 3; long dashed line, linear loess smoother, span $f = 0{\cdot}50$. Notice that the residuals from the non-linear smoother oscillate least, indicating overall a successful fit except at point 5, a clear unusual value

series with noise that includes an extreme value (at point 5), a peak (points 9–20), and an abrupt shift (between points 25 and 26). The solid line, running *medians* of length 3 (3R), comes closest to the error-free sequence (dotted line), whereas running *means* of length 3 smoother (short dashed line) is influenced strongly by point 5, lowers the peak somewhat, and suggests a more gradual shift at point 25. The loess smooth ($f = 0{\cdot}50$, long dashed line) is completely misleading. Because the examination of the residuals is a key reason to smooth, non-linear smoothers are more useful in highlighting such values in the residuals; Figure 1(*b*) shows that most of the 3R residuals are relatively flat, except for that at point 5. Since Tukey[22] advocated their use, non-linear smoothers for one-dimensional data have been studied in the literature, but not to any great extent.[32–34]

Hansen[18] implemented a non-linear smoother in two dimensions called head-banging and applied it to data on well depths in Japan, where fault lines created real features in the data that were preserved with head-banging but not with local linear or quadratic regression. The basic head-banging algorithm proceeds as follows:

1. For each point or area whose value, y_i, is to be smoothed, determine the NN nearest neighbours to $\mathbf{x}_i$.
2. From among these NN neighbours, define a set of pairs around the point/area, such that the 'triple' (pair plus target point at $\mathbf{x}_i$) are roughly collinear. (Formally, the angle formed by the two segments with $\mathbf{x}_i$ in the centre should not exceed, say $\pm 45^\circ$ from 180°. Denote this threshold ϕ, for example, $\phi = 45^\circ$.) Let NTRIP be the maximum number of such triples. If there are more than NTRIP pairs that satisfy the ϕ criterion, choose those whose angles are closest to 180°.
3. Let (a_k, b_k) denote the (higher, lower) of the two y-values in the kth pair, and let $A = \text{median}\ \{a_k\}$, $B = \text{median}\ \{b_k\}$.
4. The smoothed value is $\tilde{y}_i = \text{median}\ \{A, y_i, B\} = \text{median}\{\text{median}\{a_k\}, y_i, \text{median}\{b_k\}\}$.

Special rules apply to corner points for which there are no pairs that satisfy the ϕ criterion. These steps are either repeated until convergence (that is, no further changes in the smoothed values), or are repeated a specified number of iterations; the first criterion is used most often, since generally fewer than five iterations are required for convergence. Notice that the value y_i is altered only if y_i falls below the typical low or above the typical high of the points surrounding it. Obviously, greater smoothing is achieved as NN, NTRIP and ϕ are increased, with NN being the most influential on the smoothness of the result. Hansen (reference 35, Chapter 2) provides suggestions for choosing these parameter values based on a data set with 120 points: NN $\approx$ 20 per cent of n, the number of data points, NTRIP $\approx$ two-thirds of NN (or possibly as large as NN), and $\phi = 45^\circ$. (No such guidelines have been established for data sets as large as that described in Section 3.) An increase in NTRIP by 1 merely adds one value to the set from which the medians A and B are calculated, and so usually has little effect. However, an increase in NN by 1 can change the entire set of triples. Experience suggests that the head-banged smooth is less sensitive to small changes in its parameters than loess is to f, possibly due in part to the use of nested medians (that is, medians of medians), which results in a highly robust smoother (see repeated median regression[36]). Because mortality/incidence rates have widely different variances depending upon the populations on which they are based, Hansen–Simonson later developed a weighted head-banging algorithm that is appropriate for data whose weights can be specified.[16] The above algorithm still applies but with weighted medians replacing ordinary medians in the definitions of A, B (step 3) and $\tilde{y}_i$ (step 4).

Faced with all of these smoothers, the data analyst must choose both a smoother and the parameters that define it. Comparisons among smoothers suggest that relatively smooth

underlying surfaces subjected to Gaussian noise benefit most when linear smoothing is applied, but unusual values, sharp features or non-Gaussian noise render them less efficient than non-linear smoothers.[14,37] Parameters to control the amount of smoothing can be based on cross-validation, with some risk of undersmoothing, or a 'plug-in' method, with some risk of over-smoothing, so the optimal choice may lie somewhere in between.[28] Experimentation with head-banging parameters indicates somewhat less sensitivity in the output of head-banging with small changes to the parameters than occurs with the parameter f in loess, but definitive studies concerning optimal choices of head-banging parameters, beyond that done by Hansen[35] for $n = 120$, remain to be conducted.

4. THE DATA

We illustrate smoothing and adjustment on melanoma mortality rates in 3053 counties in the continental United States over the years 1973–1987, primarily because they suggest an obvious adjuster for purposes of evaluating the success of the methodology. The data come from the National Cancer Institute[38] separately for males and females and for Whites (including Hispanics) and non-Whites (including Asian Americans, Native Americans, African Americans and Alaskan Natives). Most of the counties in the 48 states are unambiguous; the exceptions are Virginia (where independent cities are incorporated into the surrounding counties) and Wisconsin (where Oconto and Shawano are included in Menominee county). Smoothing is applied to all 3053 counties. (The S-plus[39] mapping routines actually have 3082 counties in the continental United States, to account for situations such as St. Martin parish in Louisiana, which consists of two separate pieces, but the same rate applies to both.) Numbers of deaths and total person years are summed over the 15 year period by 18 five-year age groups: 0–4, 5–9, ..., 80–84, 85+. Adjustment for age is carried out using the 1970 total U.S. population as the standard. Assigned county locations for these rates are the longitude and latitude co-ordinates for the population centroid of the (at most) three largest places in the county.[40] Only the data for White females and White males are analysed here, since the rates among non-Whites are much lower. For purposes of assessing the magnitude of the rates, it is useful to note that the national age-adjusted rate per 100,000 population is 1·59 for White females and 2·84 for White males.

5. RESULTS: SMOOTHING THE AGE-ADJUSTED RATES

Both linear and non-linear smoothers have roles in analysing geographical trends in cancer mortality rates. Linear smoothers have the potential to produce very broad, definite patterns that give the impression of a stronger trend than actually exists; however, this feature serves to suggest possible variables for adjustment. The cost for these broad sweeping trends is a misleading impression about individual features and particular county rates. In regions where the underlying surface is truly smooth, non-linear smoothers can be iterated to yield as smooth a result, yet retain breaks in the surface if they are real, unlike linear smoothers that will be forced to smooth over them. Thus, a linear smoother may be more useful initially to identify a potential adjuster (variable for which data should be adjusted, beyond age), while a non-linear smoother can identify specific areas of high or low rates. The same argument will apply to smoothing the residuals, or rates that have been adjusted for these variables.

For purposes of comparison, the unsmoothed age-adjusted rates are compared with age-adjusted rates that are smoothed by loess ($f = 0{\cdot}33$), the age-specific rates smoother ($d_{\max} = 2{\cdot}5$,

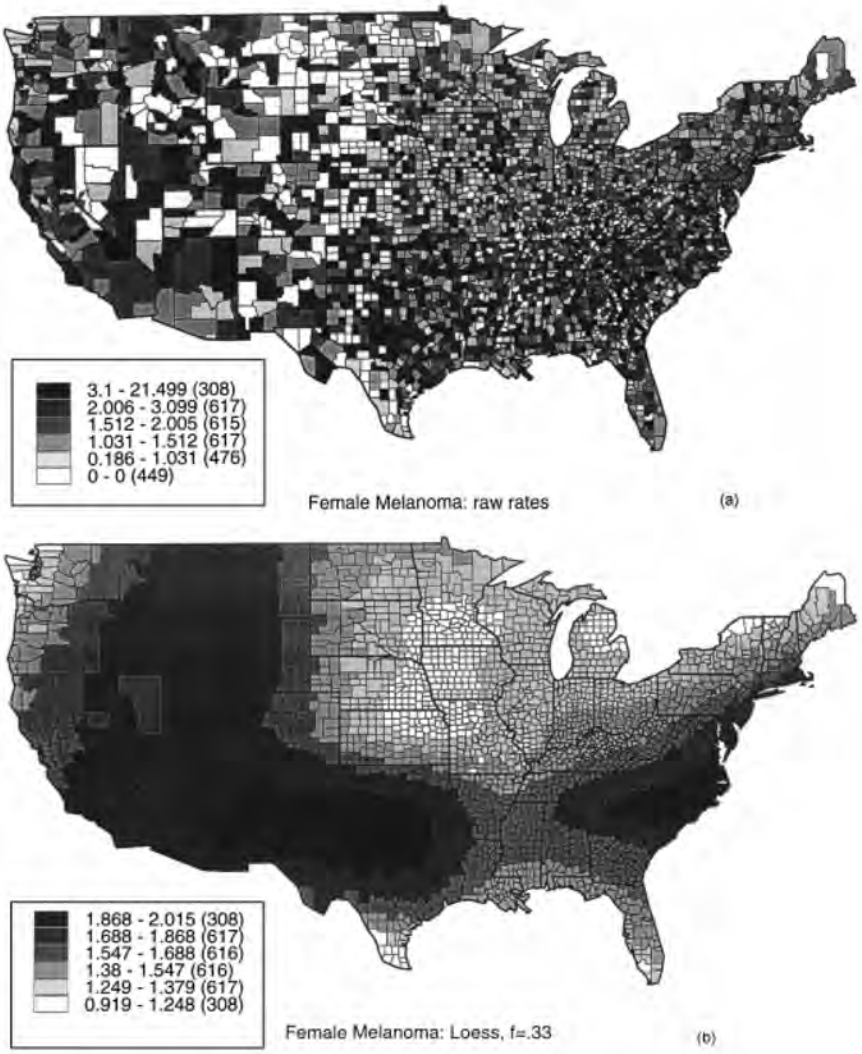


Figure 2. Female age-adjusted melanoma mortality rates, 1973–1987, in counties in the continental United States: (*a*) unsmoothed rates; (*b*) smoothed by loess, span $f = 0{\cdot}33$; (*c*) smoothed by age-specific rates smoother, weights given by equation (1); (*d*) smoothed by weighted head-banging

or 100–150 miles), and by weighted head-banging (NN = 100, NTRIP = 70, $\phi = 45^{\circ}$; large values of NN and NTRIP increase computation time). Optimal weighting of points summarized by medians are standard deviations,[41] so the weights for this smoother are the square roots of the person-years of the counties on which the rates are based. For consistency, intrinsic weights for calculating the loess smooth are the same. Neither loess nor head-banging take advantage of the particular form of the data as rates.

Figures 2 and 3 display the smoothed melanoma mortality rates for females and males, respectively. For each figure, unsmoothed rates are displayed in part (*a*), loess in (*b*), age-specific

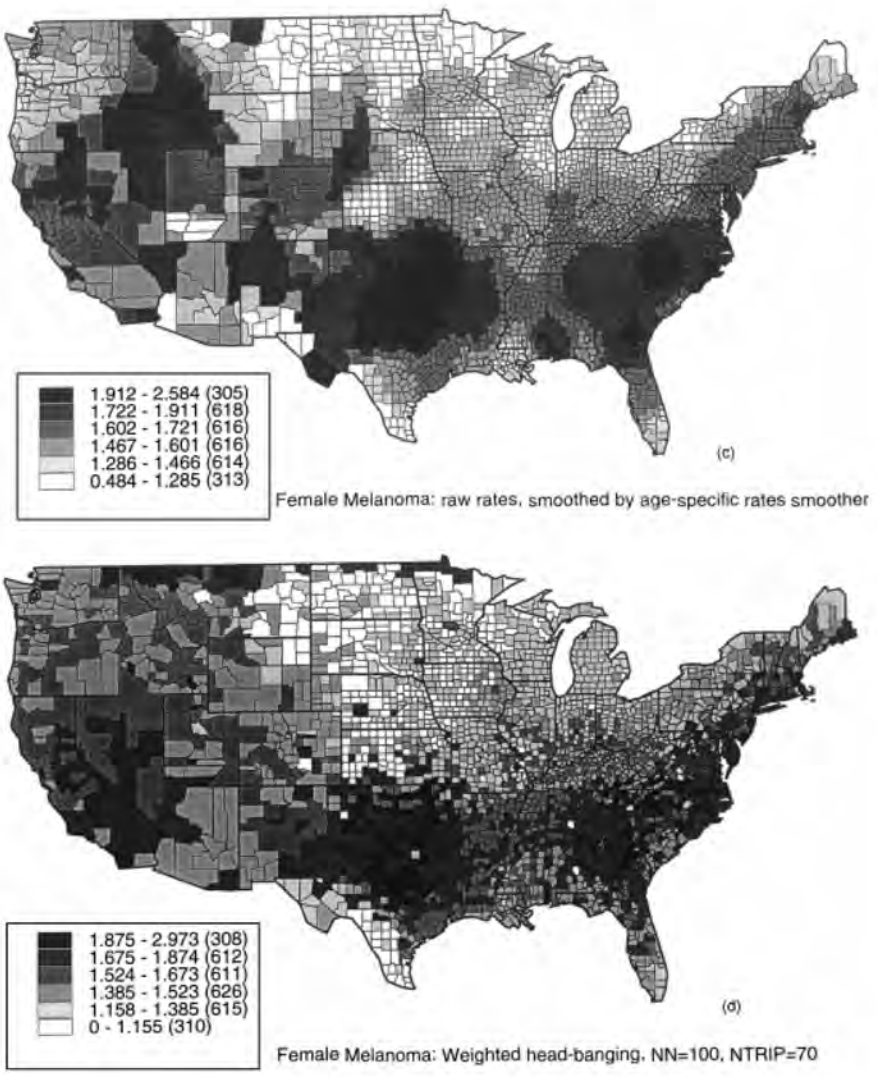


Figure 2. (*Continued*)

rates smoother in (*c*), and weighted head-banging in (*d*). Both linear smoothers yield much smoother displays than head-banging, especially loess in part (*b*); the differences among them demonstrate the trade-off between bias (highest for loess) and variance (highest for head-banging) in the smoothed rates. Head-banging requires extremely large values of the parameters NN and NTRIP to achieve smoothness comparable to loess, because counties with high weights (many people) tend to dominate the smoothed rate, regardless of how many neighbours or triples are used. All three smoothers suggest gradually decreasing rates with increasing latitudes, except possibly in the west, where the rates appear to vary little with latitude; this effect is virtually

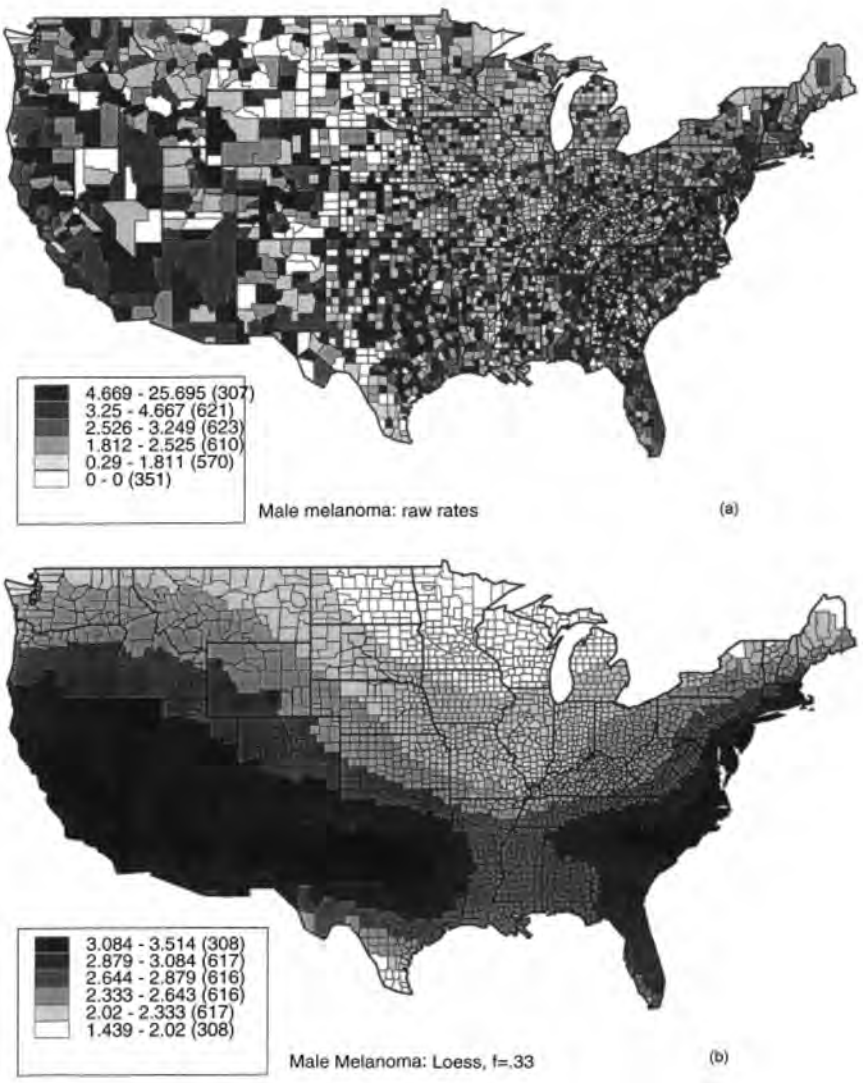


Figure 3. Male age-adjusted melanoma mortality rates, 1973–1987, in counties in the continental United States: (*a*) unsmoothed rates; (*b*) smoothed by loess, span $f = 0{\cdot}33$; (*c*) smoothed by age-specific rates smoother, weights given by equation (1); (*d*) smoothed by weighted head-banging

undetectable in the map of the unsmoothed rates in part (*a*). The latitude trend is most apparent with the linear smoothers but none the less is evident, to a lesser degree, with the non-linear smoother also. Section 7 contains detailed discussion concerning specific features of these maps; here we note only that the overly smoothed loess map leaves the impression of a smoothly varying, continuous and consistent trend with latitude, while the age-specific rates smoother demonstrates not only the latitude effect but also some individual patches of potential interest. The non-linear smoother gives a less definite impression of latitude, but also gives a more

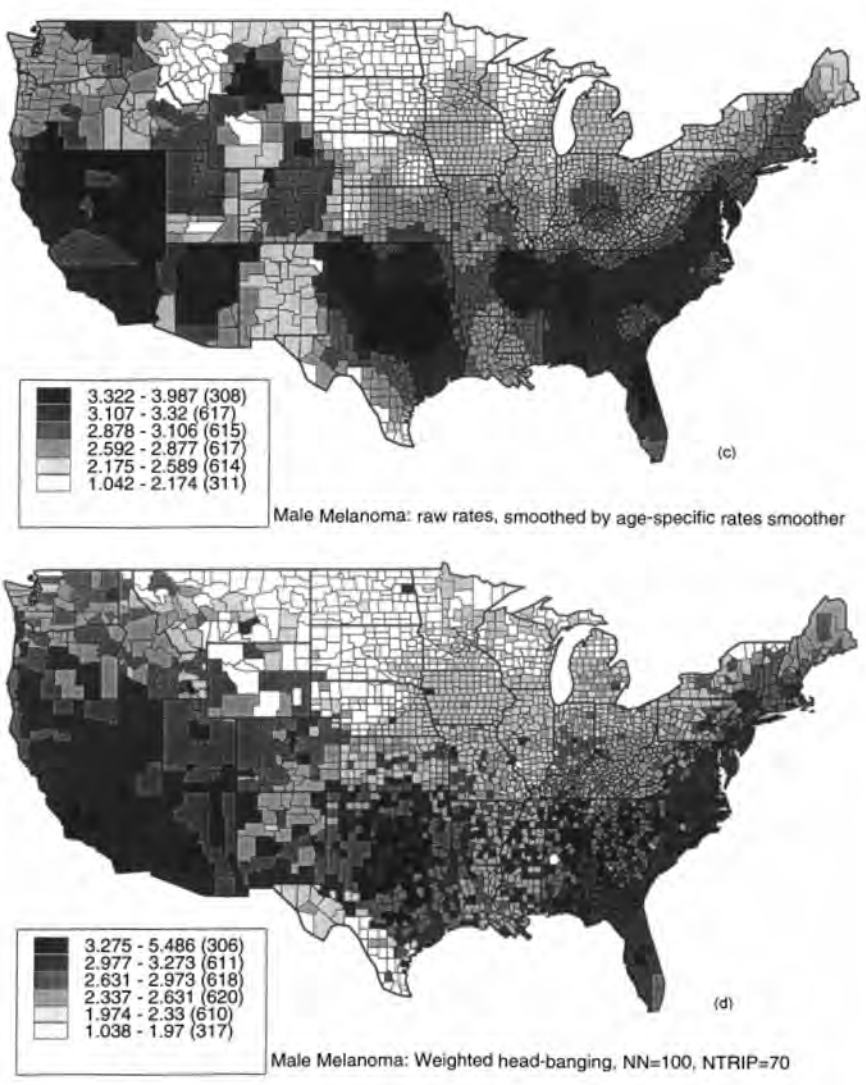


Figure 3. (*Continued*)

accurate indication of the the actual raw rates while also attenuating rates based on few person-years that are unsupported by their neighbours. Both linear smoothers in parts (*b*) and (*c*) are better for identifying the broad latitude trend, whereas the non-linear smoother is better for confirming its extent (less in the west than in the central or eastern states).

6. METHODS: ADJUSTMENT FOR LATITUDE

The smoothed rates described in the previous section suggest an obvious adjuster for melanoma rates, namely latitude. This adjuster was far from obvious in the unsmoothed rates, but it is

biologically plausible since more frequent exposure to the sun should result in rates that are higher in the south than in the north. Other variables were considered as adjusters (average January temperature, average July temperature, average number of days of sunlight), but trends based on them were not as pronounced as with latitude. We consider two ways to adjust the age-adjusted rates for latitude.

6.1. Indirect adjustment

Because of the possible difference in the effect of latitude on rates in different parts of the county, counties are stratified within region (east, central, west), according to their latitudes, into 15 strata of 1 degree latitude range for the 10 middle strata and $2-6^{\circ}$ for five end strata. Table I gives the number of person-years (rounded to the nearest 100,000) and the average stratum rate for each of the 45 region $\times$ stratum combinations. Let $s(R, L)$ denote the average stratum rate for region R ($R = 1, 2, 3$) and latitude stratum L ($L = 1, \ldots, 15$). The average stratum rate is calculated by pooling all deaths and all populations by age group across all counties in the stratum and age-adjusting accordingly, again using the 1970 U.S. total population as the standard. Because these stratum rates have variability, we smooth them using a one-dimensional non-linear smoother (the function *smooth* in S-plus) as a function of latitude. These smoothed trends, $\tilde{s}(R, L)$, shown in Figure 4, are used to adjust the county rates as follows:

$$a_i = r_{i(R,L)}/\tilde{s}(R, L) \tag{2}$$

where $r_{i(R,L)}$ denotes the ith county rate located in region R and latitude stratum L. Values of a_i in (2) that exceed 1.0 indicate rates that are higher than expected for their latitudes. Before plotting, the a_i values of (2) are smoothed by either loess ($f = 0{\cdot}33$) or by weighted head-banging as described in Section 3.

6.2. Simultaneous direct adjustment and smoothing

A parallel to the age-specific rates smoother for the adjusted rates avoids some of the roughness that appears when using the weighted head-banging smoother and utilizes the age categories for each rate. Above, we calculated a stratum rate by age-adjusting the pooled deaths and populations within a stratum and smoothed the ratio (observed rate/expected rate). Here, we propose to first smooth the age-specific trends as a function of latitude and then age-adjust the smoothed trends. This procedure permits the application of the age-specific rates smoother so that the resulting smoothed rates may suggest further variables for adjustment. Let $s_m(R, L)$ be the age-specific rate in age group m, $m = 1, \ldots, 18$, by pooling all deaths and person-years for this age group in counties from region R and latitude stratum L. Several rates are quite variable due to small populations in each age group $\times$ region $\times$ stratum combination, so, instead of using $s_m(R, L)$ directly, we smooth them as a function of latitude L, now for each region (3) $\times$ age-group (18) category, and denote the smoothed trend $\tilde{s}_m(R, L)$. (The first four age groups in each region have smoothed trends $\tilde{s}_m(R, L) = 0$, $m = 1, \ldots, 4$, $L = 1, \ldots, 15$.) We then adjust the age-specific rates for $\tilde{s}_m(R, L)$ followed by an adjustment for age, that is

$$\begin{aligned} a_{im} &= r_{i(R,L)m}/\tilde{s}_m(R, L) \quad \tilde{s}_m(R, L) > 0 \\ &= 0 \qquad\qquad\qquad\qquad\quad \tilde{s}_m(R, L) = 0 \end{aligned}$$

Table I. Stratum rates for 45 region × latitude strata (person-years, per 100,000, in parentheses)

Latitude range	East*		Central†		West‡	
Females						
24–30	1·527	(560)	1·486	(458)	0·000	(0)
30–32	1·871	(184)	1·774	(366)	1·630	(8)
32–34	1·819	(378)	1·888	(435)	1·671	(503)
34–35	1·806	(255)	1·914	(117)	1·587	(604)
35–36	2·033	(398)	1·946	(166)	1·787	(94)
36–37	1·886	(324)	1·916	(127)	1·592	(131)
37–38	1·631	(288)	1·463	(122)	1·632	(354)
38–39	1·647	(461)	1·461	(213)	1·815	(167)
39–40	1·604	(946)	1·412	(135)	1·664	(169)
40–41	1·642	(1908)	1·462	(71)	1·616	(142)
41–42	1·425	(1522)	1·348	(151)	1·924	(35)
42–43	1·508	(1182)	1·321	(102)	1·540	(47)
43–44	1·234	(440)	1·429	(62)	1·896	(49)
44–46	1·364	(347)	1·307	(254)	1·419	(196)
46–50	1·282	(33)	1·156	(102)	1·508	(323)
Males						
24–30	3·279	(518)	2·852	(449)	0·000	(0)
30–32	3·254	(178)	3·053	(362)	2·588	(7)
32–34	3·070	(364)	3·299	(416)	3·297	(494)
34–35	3·340	(246)	2·951	(112)	3·159	(585)
35–36	3·388	(379)	3·196	(158)	2·728	(92)
36–37	2·886	(313)	2·940	(121)	2·844	(131)
37–38	2·751	(277)	2·545	(116)	3·135	(345)
38–39	2·723	(438)	2·790	(200)	3·255	(164)
39–40	2·989	(888)	2·532	(127)	3·134	(166)
40–41	2·872	(1759)	2·099	(67)	3·150	(141)
41–42	2·536	(1429)	2·210	(143)	3·182	(36)
42–43	2·645	(1109)	2·162	(98)	2·960	(47)
43–44	2·332	(418)	2·255	(60)	2·709	(49)
44–46	2·086	(333)	2·107	(244)	2·643	(191)
46–50	2·316	(34)	1·684	(102)	2·580	(319)

* East includes counties in Maine, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island, New York, Pennsylvania, New Jersey, Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida, Alabama, Tennessee, Kentucky, Ohio, Indiana, Illinois, Michigan, Wisconsin

† Central includes counties in Minnesota, Iowa, Missouri, Arkansas, Mississippi, Louisiana, Texas, Oklahoma, Kansas, Nebraska, North Dakota, South Dakota

‡ West includes counties in Montana, Wyoming, Colorado, New Mexico, Arizona, Utah, Idaho, Washington, Oregon, Nevada, California

and the smoothed adjusted rate is

$$\tilde{a}_i = \sum_{m=1}^{18} \pi_m a_{im}^*$$

where

$$a_{im}^* = \sum_{j=1}^{N} w_{ij} a_{im}$$

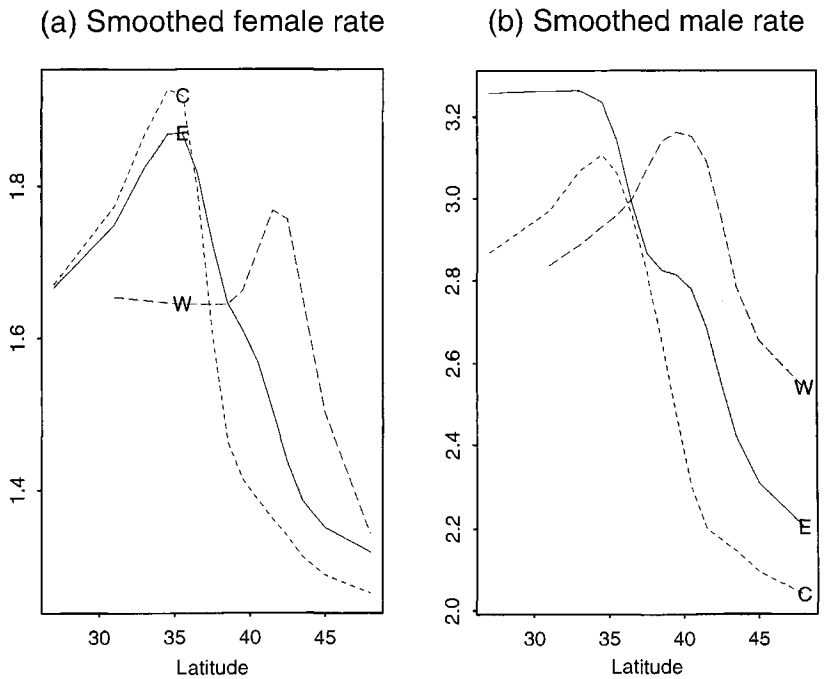


Figure 4. Smoothed melanoma rates as a function of latitude: solid (E), eastern region; dot (C), central region; dash (W), western region; (*a*) females; (*b*) males

where w_{ij} are the same intercounty distance weights multiplied by the square root of the person-years in county j, as in (1).

Figures 5 (females) and 6 (males) display the adjusted rates, again with unsmoothed in part (*a*) and three smoothed versions in part (*b*) [loess smooth of a_i in equation (2), $f = 0{\cdot}33$], (*c*) [simultaneous adjustment for age and latitude, smoothed per equation (3)), and (*d*) (headbanging smooth of a_i in equation (2), NN = 100, NTRIP = 70, $\phi = 45^\circ$).

7. DISCUSSION: RESULTS OF SMOOTHING THE LATITUDE- AND AGE-ADJUSTED RATES

The displays in Figures 2,3, 5 and 6 involve county shadings, the densities of which are indicative of the size of the rate. The smoothed county values are categorized into the highest and lowest 10 per cent and the four middle 20 per cent ranges, as nearly as the numbers of counties in each region will allow. For an extensive discussion of the cognitive perception tasks associated with map reading, see Pickle and Herrmann.[42]

Except for the western region, the age-adjusted rates in Figures 2 and 3 increase with increasing latitude for both White males and White females. This pattern is not obvious from the maps of the unsmoothed rates (part (*a*)), confirming the need for smoothing even when a plausible adjuster with biological significance is present. At the opposite extreme, loess (part (*b*)) smooths the rates

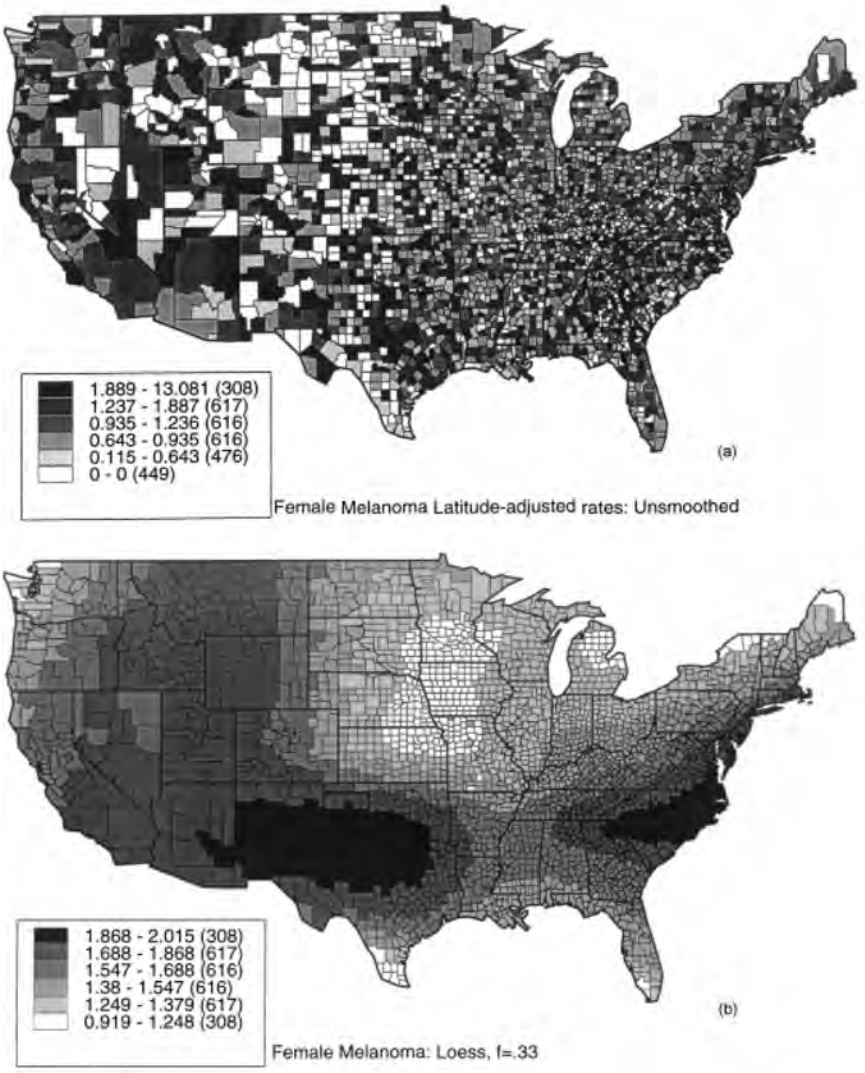
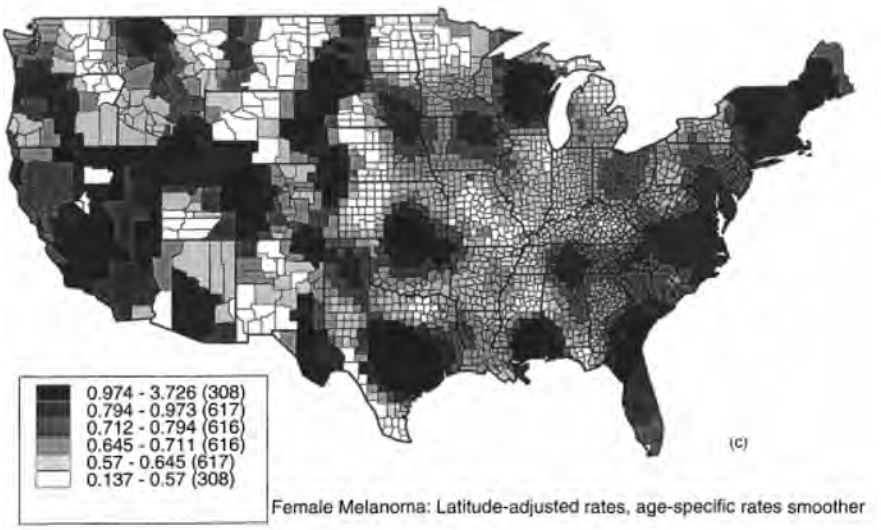


Figure 5. Female age- and latitude-adjusted melanoma mortality rates, 1973–1987, in counties in the continental United States. (*a*) unsmoothed rates; (*b*) smoothed by loess, span $f = 0{\cdot}33$; (*c*) smoothed by age-specific rates smoother, using equation (3); (*d*) smoothed values of equation (2) by weighted head-banging

almost to excess, even with a span ($f = 0{\cdot}33$) that is less than half the value of the default span in S-plus ($f = 0{\cdot}75$). The age-specific rates smoother, while also quite smooth, provides more local features to the overall latitude trend and draws attention to parts of the country whose elevated rates might not be so noticeable otherwise (central Texas and the Carolinas in females; Alabama, central Tennessee, and central Texas in males), while reducing attention to rates that are high only because of one or two deaths in sparsely populated counties (for example, southwestern Colorado where the most darkly shaded counties correspond to counties of less than 70,000



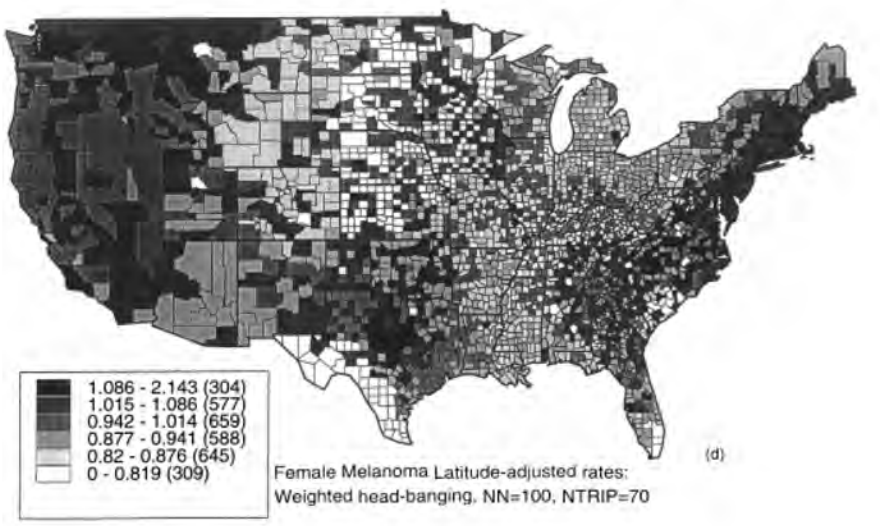


Figure 5. (*Continued*)

person years for the 15 years 1973–1987, such as Hillsdale, with a 1980 population of only 408). However, this smoother also has a tendency to create patches out of single highly populated, high rate counties in the area, such as in Nebraska for White females (due to Hall and Lincoln counties, with eight and nine deaths, and 366,832 and 260,393 person years, respectively), and, for White males, in Nevada (due to Clark, Washoe, and Carson City counties, which together constitute 85 per cent of the total deaths and person-years in this state), in south central Montana (due to Yellowstone county, with 28 deaths and 776,376 person years, surrounded by counties of no more than 50,000 person years), and in northern California (where relatively high rates occur in Inyo (7·01), Tuolomne (5·75) and Siskiyou (5·66) counties). Head-banging is a reasonable

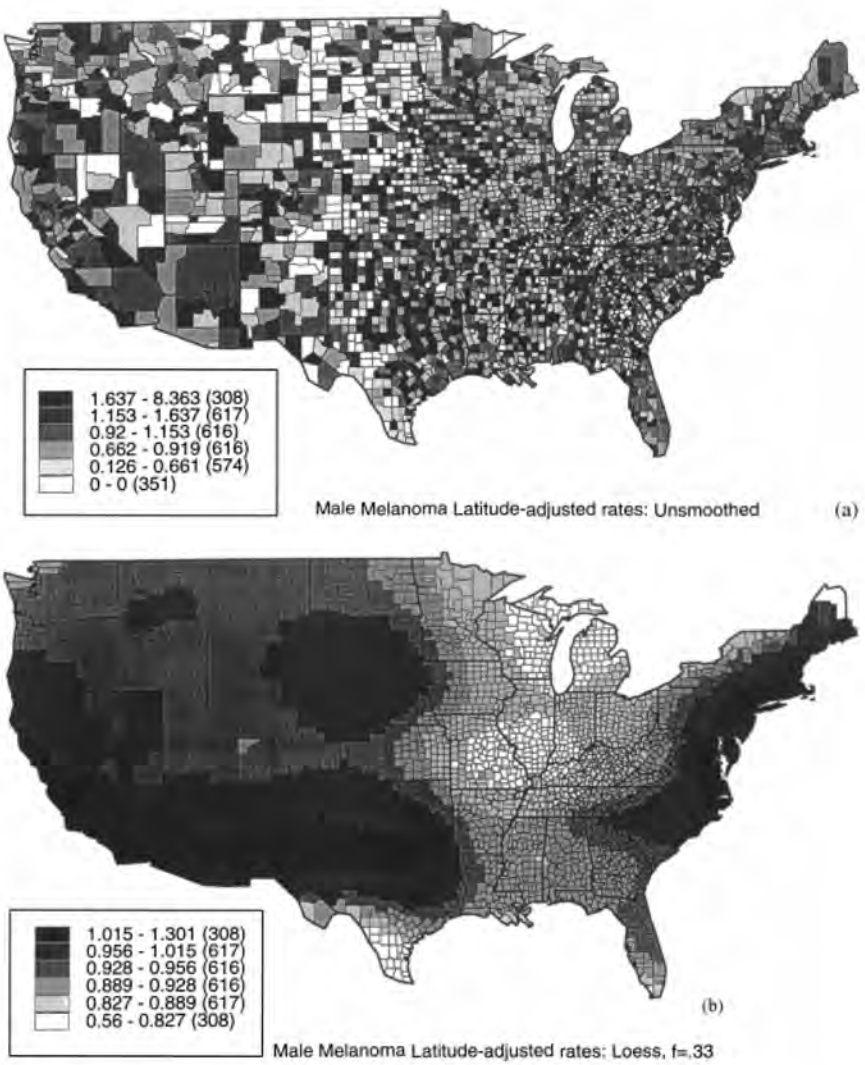


Figure 6. Male age- and latitude-adjusted melanoma mortality rates, 1973–1987, in counties in the continental United States: (*a*) unsmoothed rates; (*b*) smoothed by loess, span $f = 0{\cdot}33$; (*c*) smoothed by age-specific rates smoother, using equation (3); (*d*) smoothed values of equation (2) by weighted head-banging

compromise between the unsmoothed rates in part (*a*) and the oversmoothed loess rates in part (*b*); while not as smooth as the age-specific rates smoother in part (*c*), the map nonetheless provides a definite indication of a trend with latitude, but reflects greater accuracy in the true rates (for example, high rates among White females in the northernmost counties of Montana, North Dakota and Minnesota, in Colorado's Arapahoe county adjacent to Denver, and in Nebraska's Lincoln county, home of North Platte, the state's third most populous city in 1980; high rates among White males in Alabama, central Tennessee, the Carolinas, central Texas, Yellowstone in Montana, and Boulder and El Paso counties in Colorado).

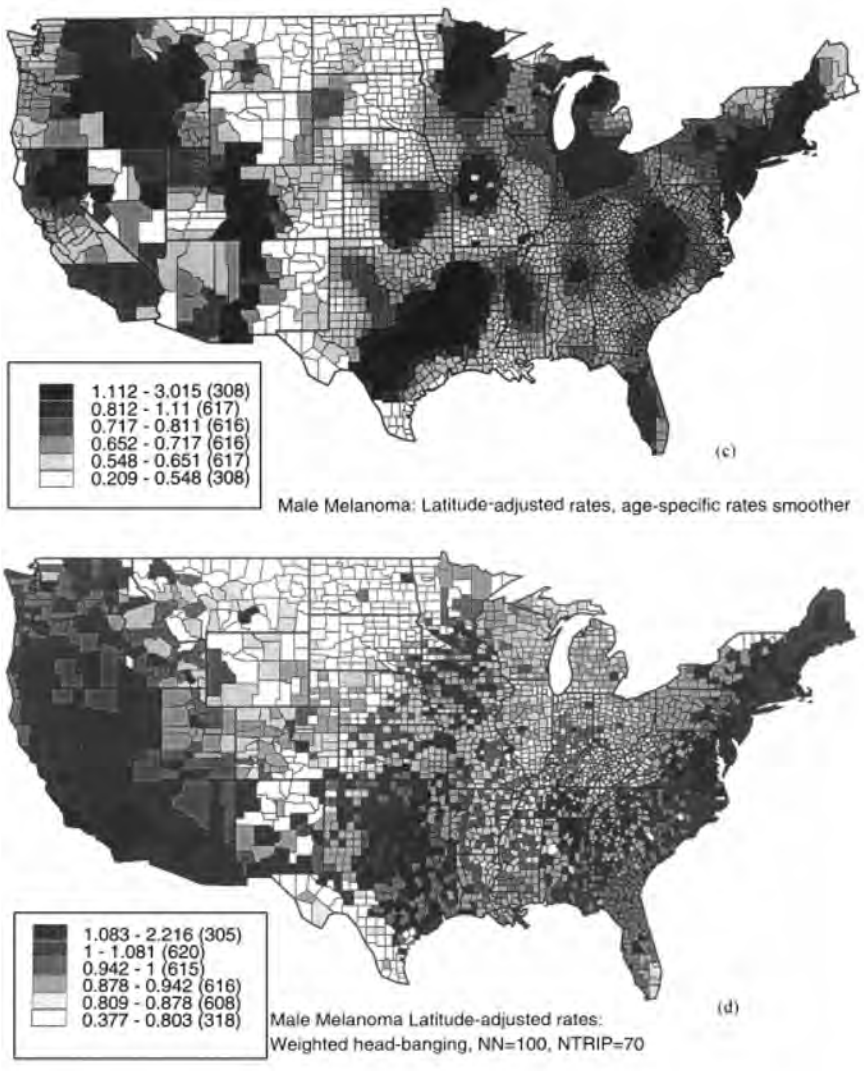


Figure 6. (*Continued*)

Once the rates have been adjusted for latitude, different patterns emerge. The median rate in each set of rates is now 1, as is expected from any successful adjustment (observed rate/fitted latitude rate). As with the unadjusted rates, clear patterns are difficult to decipher without the benefit of smoothing (part (*a*)). Loess gives the impression of very definite patterns (part (*b*)): excessive high rates in the northwestern plain states among females, high rates throughout the southwest and along the California coast among males, and high rates along the east coast and low rates in Illinois and Montana for both males and females. The age-specific rates smoother and head-banging in parts (*c*) and (*d*), respectively, fail to confirm the loess-indicated high female rates throughout Idaho, Montana, and Wyoming, except perhaps in a few northernmost counties

bordering Canada. The rates along the east coast for both males and females are still relatively high given their northernmost latitudes, especially in New England; one might attribute this finding to the enticing proximity of the Atlantic coast. By way of contrast, many southern counties now have lower rates, once they have been adjusted for latitude, except for a patch of counties in central Texas in both males and females. Such patches created by the age-specific rates smoother does serve to draw one's attention to an area of potential interest, but then such patches really must be confirmed by referring to either the map of the non-linearly smoothed rates, where patterns may not be so evident (see eastern Iowa, southwestern Colorado) but which better reflect the underlying rates (again, less bias/higher variance, than with the age-specific smoother), or to the raw data directly, by checking to see if the high rates are based on sufficiently large counts of deaths and persons. Among males, low rates are evident in the northwestern plain states (except noticeably high rates in both Grand Forks, North Dakota, and Yellowstone, Montana), high rates appear along the east coast upwards into New England and New York, and scattered high rates in Iowa and southern Minnesota. So the broad-brush linear smoothers can be useful, so long as they impose only moderate bias (as is true with the age-specific rates smoother) and so long as the patterns observed from them can be confirmed with a less biased smoother (such as head-banging).

8. SUMMARY

In each comparison of displays of the rates, patterns in the unsmoothed rates are almost indecipherable, while those in the loess smoothed rates may be too broad to be entirely valid. Between these two extremes, the age-specific rates smoother, while also linear and thus somewhat broad-brush, succeeds in suggesting patterns and still retains some specific regional patterns of high and low rates. The greatest danger lies in its tendency to create patches out of one or two high-rate, highly populated counties. Non-linear smoothers require larger values of smoother parameters and more iteration to achieve comparable smoothness, but in general they are more successful at recovering sharp features such as isolated high-rate counties or sudden shifts in levels of rates.

These results demonstrate an important principle of smoothing: no one smoother is likely to achieve all purposes at all times. A comparison of the displays using the linear age-specific rates smoother (1) and the non-linear head-banging smoother shows that the former highlights very clearly possible trends, but often with a substantial amount of bias. Conversely, non-linear smoothers tend to show somewhat greater variability (less smooth) but are also less biased and hence more faithful to particular, sometimes sharp, features in the data. Thus, both smoothers are useful, both serve important purposes, and both provide insights that can be studied further with other methods.

Melanoma rates suggest a biologically sensible adjuster and thus is an ideal data set on which to evaluate the smoothing methodology proposed here. However, even this 'obvious' adjuster was less than obvious in the map of the unsmoothed rates, thus demonstrating the value of smoothing these data. The methods in this paper demonstrate an approach to analysing data of this sort, particularly in those circumstances where the adjuster variable was less obvious: use the broad-brush age-specific rates smoother to suggest the adjuster, confirm the patterns with a non-linear smoother such as head-banging, and repeat the process on the rates adjusted for the suggested variable.

Finally, even when the expected effects are not evident (for example, the lack of a latitude effect in the west), such an observation can be useful in its own right. At the symposium where these maps were first shown, one participant remarked to me later that the migration within the western region in both directions (Seattle to San Diego for retirees; San Diego to Seattle for job seekers) may prevent a consistent stable population in any one area that would be necessary to see a latitude effect. This remark suggests that, to confirm a latitude effect, some detailed information concerning individual exposures would be necessary and would have to be conducted by studies other than geographical epidemiological ones (for example, via case-control studies). The geographic analysis here once again is valuable for suggesting alternative hypotheses to be confirmed by other studies.

ACKNOWLEDGEMENTS

The author thanks Dr. Kenneth H. Falter, Guest Editor, and the referees, for many helpful comments on an earlier version of this paper. Support for this research from the National Science Foundation is gratefully acknowledged, through grant number DMS 95-10435, awarded to the University of Colorado at Denver.

REFERENCES

1. Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B.,Creason, J. P. and Pellom, A. C. 'Empirical Bayes procedures for stabilizing maps of U. S. cancer mortality rates', *Journal of the American Statistical Association*, **84**, 637–650 (1989).
2. Marshall, R. J. 'Mapping disease and mortality rates using empirical Bayes estimators', *Applied Statistics*, **40**, 283–294 (1991).
3. Clayton, D. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping,' *Biometrics*, **43**, 671–681 (1987).
4. Kaldor, J. and Clayton, D. 'Role of advanced statistical techniques in cancer mapping', in Boyle, P., Muir, C. S. and Grundmann, E. (eds), *Recent Results in Cancer Research* 114: *Cancer Mapping*, Springer-Verlag, New York, 1989, pp. 87–98.
5. Cislaghi, C., Biggeri, A., Braga, M., Lagazio, C. and Marchi, M. 'Exploratory tools for disease mapping in geographic epidemiology', *Statistics in Medicine* **14**, 2363–2382 (1995).
6. Besag J., York, J. and Mollié, A. 'Bayesian image restoration with two applications inspatial statistics (with discussion)', *Annals of the Institute of Statistical Mathematics*, **43**, 1–59 (1991).
7. Cressie, N. 'Smoothing regional maps using empirical Bayes predictors', *Geographic Analysis*, **24**, 75–95 (1992).
8. Devine, O. J., Louis, T. A. and Halloran, M. E. 'Empirical Bayes estimators for spatially correlated incidence rates', *Statistics in Medicine*, **5**, 381–398 (1994).
9. Devine, O. J., Louis, T. A. and Halloran, M. E. 'Identifying areas with elevated disease incidence rates using empirical Bayes estimators', *Geographic Analysis* **28**, 187–199 (1996).
10. Devine, O. J., Louis, T. A. and Halloran, M. E. 'Empirical Bayes methods for stabilizing incidence rates before mapping,' *Epidemiology*, **5**, 622–630 (1994).
11. Bloomfield, P. 'Analysis of cancer mortality in Texas', *Proceedings of the* 1976 *Workshop on Automated Cartography*, DHEW Publication No. (PHS) 79–1254, 1979, pp. 27–33.
12. Carr, D. B. and Pickle, L. W. 'Topics in scientific visualization: Plot production issues and details', *Statistical Computing and Statistical Graphics Newsletter*, **4**, 16–20 (1993).
13. Cressie, N. and Read, T. A. C. 'Spatial data analysis of regional counts', *Biometrical Journal* **31**, 699–719 (1989).
14. Kafadar, K. 'Choosing among two–dimensional smoothers in practice', *Computational Statistics and Data Analysis*, **18**, 419–439 (1994).
15. Kafadar, K. 'Smoothing geographical data, particularly rates of disease', *Statistics in Medicine*, **15**, 2539–2560 (1996).

16. Mungiole, M., Pickle, L. W., Simonson, K. H. 'Application of a weighted head-banging algorithm to mortality data maps', *Statistics in Medicine*, **18**, 000–000 (1999).
17. Cleveland, W. S. and Devlin, S. J. 'Locally weighted regression: an approach to regression analysis bylocal fitting', *Journal of the American Statistical Association*, **83**, 596–610 (1988).
18. Hansen, K. M. 'Head-banging: robust smoothing in the plane', *IEEE Transactions on Geoscience and Remote Sensing*, **29**, 369–378 (1991).
19. Nychka, D., Gray, G., Haaland, P., Martin, D. and O'Connell, M. 'Syringe grading based on extracted features from high dimensional friction data', *Journal of the American Statistical Association*, **90**, 1171–1178 (1995).
20. Tukey, J. W. 'Ecological exploration: a series of reports based on a visit to the National Cancer Institute, Department of Statistics, Princeton University, 1993.
21. Tukey, J. W. 'Statistical mapping: what should not be plotted', in *Proceedings of the 1976 Workshop on Automated Cartography*, DHEW Publication No. (PHS) 79–1254, 1979, pp. 18–26. Reprinted in Cleveland, W. S. (ed.), *The Collected Works of John W. Tukey*, *Volume V*: *Graphics*, *1965–1985*, Wadsworth, Belmont, California, 1988, pp. 109–121.
22. Tukey, J. W. *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts, 1977.
23. Ripley, B. *Spatial Statistics*, Wiley, New York, 1981.
24. Wahba, G. *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.
25. Yang, C. J. and Huang, T. S. 'The effect of median filtering on edge location estimation', *Computer Graphics and Image Processing*, **15**, 224–245 (1981).
26. Narendra, P. M. 'A separable median filter for image noise smoothing', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **3**, 20–29 (1981).
27. Tukey, P. A. and Tukey, J. W. 'Graphic display of data sets in 3 or more dimensions', in Barnett, V. (ed.), *Interpreting Multivariate Data*, Wiley, Chichester, 1981, pp. 189–275. Reprinted in Cleveland, W. S. (ed.), *The Collected Works of John W. Tukey*, *Volume V*: *Graphics*, *1965–1985*, Wadsworth, Belmont, CA, 1988, pp. 188–288.
28. Simonoff, J. *Smoothing Methods in Statistics*, Springer-Verlag, New York, 1996.
29. Chen, L. 'Multivariate regressions splines', *Computational Statistics and Data Analysis*, **26**, 71–82 (1997).
30. Hastie, T. and Loader, C. 'Local regression: Automatic kernel carpentry (with discussion)', *Statistical Science*, **8**, 120–143 (1993).
31. Mosteller, F. and Tukey, J. W. *Data Analysis and Regression*: *A Second Course in Statistics*, Addison-Wesley, Reading, Massachusetts, 1977.
32. Mallows, C. L. 'Some theory of nonlinear smoothers', *Annals of Statistics*, **8**, 695–715 (1980).
33. Velleman, P. F. 'Definition and comparison of robust nonlinear data smoothing algorithms', *Journal of the American Statistical Association*, **75**, 609–615 (1980).
34. Gebski, V. and McNeil, D. 'A refined method of robust smoothing', *Journal of the American Statistical Association*, **79**, 616–623 (1984).
35. Hansen, K. M. 'Some statistical problems in geophysics and structural geology', PhD Dissertation, Department of Statistics, Princeton University, 1989.
36. Siegel, A. F. 'Robust regression using repeated medians', *Biometrika*, **69**, 242–244 (1982).
37. Emsermann, M. and Kafadar, K. 'Smoothing non-Gaussian data in two dimensions', *Computing Science and Statistics* **29**, 178–181 (1998).
38. Pommerenke, F. A. and Srivastava, S. 'State cancer control map and data program targeting cancer control at the local level', in Greenwald, P., Kramer, B. S. and Weed, D. L. (eds), *Cancer Prevention and Control*, Marcel Dekker, New York, 1995, pp. 771–775.
39. *S-plus User's Manual*, *Version* 3. 3 *for Sun SPARC*, *SunOS* 5. 3, Mathsoft, Inc., Seattle, Washington, 1995.
40. Goodall, C. R., Kafadar, K. and Tukey, J. W. 'Comparing and computing some measures of urbanicity', *American Statistician*, **52**, 101–110 (1998).
41. Tukey, J. W. *Notes to 1976 Statistics 411*, Department of Statistics, Princeton University, 1976.
42. Pickle, L. W. and Herrmann, D. (eds). 'Cognitive aspects of statistical mapping', National Center for Health Statistics Working Paper Series Report No. 18, Hyattsville, Maryland, 1995.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3189–3200 (1999)

MODEL-BASED SMALL AREA ESTIMATES OF OVERWEIGHT PREVALENCE USING SAMPLE SELECTION ADJUSTMENT

DONALD MALEC[1]*, WILLIAM W. DAVIS[2] AND XIN CAO[3]

[1]*U.S. Bureau of the Census, Statistical Research Division, room 3132-4, Washington, D.C. 20233, U.S.A.*
[2]*Westat, 1650 Research Blvd., Rockville, MD 20850-3129, U.S.A.*
[3]*IndexCo, LLC, One Corporate Center, Hartford, CT 06103, U.S.A.*

SUMMARY

Using a hierarchical model with an adjustment for sample selection, we estimate the overweight prevalence for adults, by states, using data from the Third National Health and Nutrition Examination Survey (NHANES III). A two-stage hierarchical model was selected to account for geographic variability of outcomes and to model possible overdispersion of estimates due to cluster sampling. We compare our model-based estimates with design-based estimates at the national level and obtain excellent agreement. We also provide a check of our model at the state level by comparing estimates with design-based and synthetic estimates. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

There is a continuing need to assess health status, health practices and health resources at both the national and subnational level. Estimates of these health items help determine the demand for health care and the access individuals have to it. Although the NCHS personal interview surveys can provide much of this information at the national level, little can be provided for states and counties because of excessive field costs. Making design-based state estimates from the current NHANES III is problematic for several reasons. The main reason is that no sample, at all, is selected in many states. For states with a large sample size, design-based state estimates can still be of low quality due to the high degree of geographic clustering of the sample into primary sampling units (PSUs). The need for subnational health statistics exists, however, because health and health care characteristics are known to vary geographically. Also, health care planning often takes place at the state and local level.

One alternative approach for producing subnational estimates has been to, effectively, increase the sample size by utilizing models defined across the subnational areas.[1] A challenge has been to use models realistic enough to produce accurate estimates. Towards this end, hierarchical models (models which include geographic variation among rates and can account for overdispersion due

* Correspondence to: Donald Malec, U.S. Bureau of the Census, Statistical Research Division, room 3132-4, Washington, D.C. 20233, U.S.A. E-mail: donald.j.malec@ccmail.census.gov

CCC 0277–6715/99/233189–12$17.50

to cluster sampling) have been adapted to small area estimation.[2] With the availability of Markov chain Monte Carlo (MCMC) methods,[3] estimates (and precision estimates) can be made that account for all model errors. Given current resources, model-based estimates can be made for subnational levels. At a minimum, a measure of the geographic variability of health characteristics can be determined and used to make decisions about which health characteristics to measure in future surveys of small areas.

In this paper we present a methodology for making subnational estimates which extends the hierarchical model of Malec *et al.*[2] by including an oversampling (that is, unequal selection probability) component in the likelihood. We illustrate the methodology by estimating the adult overweight prevalence by state using data from the Third National Health and Nutrition Examination Survey (NHANES III). The methodology is general and is especially useful for producing subnational estimates which, at a national level, should closely agree with design-based estimates.

1.1. NHANES III: Survey Design

NHANES III is a stratified, multi-level, clustered, personal interview survey of households[4] that was conducted in two phases: during the years 1988–1991 and 1991–1994. Sampled persons provide health and dietary information through a questionnaire and also through a physical exam. Persons were selected to represent the civilian, non-institutional population of the United States and provide national characteristics and nutrition status for the entire population and age, race and ethnic subgroups. NHANES III was designed to oversample the two largest minority groups of the U.S. population, Blacks and Mexican Americans.

1.2. Overweight prevalence in U.S.

Overweight is associated with a number of adverse health outcomes including mortality[5] and has become an increasing problem for adults in the United States.[6,7] Overweight is typically defined in terms of body mass index (BMI) which is defined by

$$\text{BMI} = \text{weight}/\text{height}^2. \tag{1}$$

Expressing BMI in the units kg/m^2, overweight is defined as $\geqslant 27.8$ for adult men and $\geqslant 27.3$ for adult women. These are the gender-specific 85th percentiles of BMI for men and women aged 20 to 29 from NHANES II (1976–1980). We refer to an adult with a BMI value below the gender specific threshold as 'normal'.

1.3. Description of the Finite Population

Let Y_{tidj} denote the overweight status for the jth individual, in demographic category d, in county i, during phase t of sampling. In particular, $Y_{tidj} = 1$ denotes overweight status as determined by BMI and $Y_{tidj} = 0$ denotes normal. Let N_{id} be the total number of individuals in demographic group d and county i as measured in the 1990 census. Here, a demographic group d describes a type of person. Specifically, d is defined as a classification (g, r, a), where g denotes gender, r denotes race/ethnicity (non-Hispanic White, non-Hispanic Black, and Mexican-American) and 'a' denotes an age category (20–24, 25–29, ..., 75–79, 80+). The six cross-classifications defined by crossing gender with race/ethnicity are important later, and each will be denoted by $c = (g, r)$.

Of interest are estimates of the finite population mean for individual characteristics defined by groupings of 'd' for local areas defined by county groupings. That is

$$\theta_{LD} = \frac{\sum_{i\in L}\sum_{d\in D}\sum_{j=1}^{N_{id}} Y_{idj}}{\sum_{i\in L}\sum_{d\in D} N_{id}} \tag{2}$$

where L indexes a particular collection of counties (for example, all counties in a specific state), and D is the set of specific subgroups of interest (for example, all females regardless of age or race). Here, Y_{idj} denotes overweight status at census day 1990.

2. ESTIMATION METHODOLOGY WITHOUT OVERSAMPLING

It is well known that ignoring the sample design can cause a selection bias and lead to erroneous conclusions.[8] A detailed discussion of the use of models to eliminate selection bias (of any kind) can be found in chapter seven of Gelman *et al.*[9] As described therein, a non-ignorable design is a design that results in selection bias while an ignorable design does not. A non-ignorable design can be made ignorable by adding appropriate design variables to the model (that is, the selection bias can be eliminated by including variables in the model that account for it). Realistically, incorporating appropriate sample design information into a model can either minimize or eliminate problems associated with selection bias.

This basic approach was adhered to by Malec *et al.*[2] in estimating small areas using the National Health Interview Survey. There, a two-stage hierarchical model was employed to account for overdispersion due to correlation within primary sampling units and a stepwise variable selection procedure was employed on the socio-economic variables used as stratifiers in the design. In addition, missed components of variance in the model were checked using cross-validation methods.

In Section 2.1, we specify a general method for estimating prevalence at a subnational level, similar to Malec *et al.*[2] This method is based on a population model that is appropriate when the sample selection is ignorable, given the two-stage model including covariates. In Section 3.1 we extend this model and estimation method to include a sample design that is non-ignorable, given the two-stage model including covariates. In Section 5 we employ data-based methods to check the adequacy of using this method.

2.1. The Population Model

A two-stage hierarchical model is used to describe individual and county variation. Conditional on the parameters p_{tid}, the Y_{tidj} are assumed to be independent Bernoulli random variables with

$$\Pr(Y_{tidj}\,|\,p_{tid}) = p_{tid}^{Y_{tidj}}(1-p_{tid})^{1-Y_{tidj}}. \tag{3}$$

Although we allowed county covariates related to the stratifiers to be candidates for model selection, these covariates did not explain much of the variation. We found the following specification of the model adequate for overweight status:

$$\text{logit}\{p_{tid}\} = \alpha_{td} + \beta_{ic} \tag{4}$$

where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_i$ denote the fixed and random effects, respectively. The components of $\boldsymbol{\beta}_i^{\mathrm{T}} = (\beta_{i1}, \ldots, \beta_{i6})$ represent the six race/gender groups. The density of $\boldsymbol{\beta}_i$, $f(\boldsymbol{\beta}_i\,|\,\Gamma)$, is a

multivariate Gaussian density with mean vector zero and dispersion matrix, Γ, that is

$$f(\boldsymbol{\beta}_i \mid \Gamma) \propto |\Gamma|^{-1/2} \mathrm{e}^{-1/2\boldsymbol{\beta}_i^{\mathrm{T}}\Gamma^{-1}\boldsymbol{\beta}_i}. \tag{5}$$

The likelihood L_{I} of $\boldsymbol{\alpha}$, $\{\boldsymbol{\beta}_i\}$, and Γ is proportional to the product of terms in (3) and (5) corresponding to sampled individuals and counties:

$$\begin{aligned} L_I(\{\boldsymbol{\beta}_i\}_{i\in s}, \boldsymbol{\alpha}, \Gamma) &\propto \prod_{tidj\in s} \Pr(Y_{tidj} \mid p_{tid}) \times \prod_{i\in s} f(\boldsymbol{\beta}_i \mid \Gamma) \\ &= \prod_{i\in s}\left[\prod_t \prod_d p_{tid}^{m_{tid}}(1-p_{tid})^{n_{tid}-m_{tid}}\right] |\Gamma|^{-1/2} \mathrm{e}^{-1/2\boldsymbol{\beta}_i^{\mathrm{T}}\Gamma^{-1}\boldsymbol{\beta}_i} \end{aligned} \tag{6}$$

where $tidj \in s$ denotes the set of all individuals in sample and $i \in s$ denotes the set of counties that contain sampled individuals. Also, n_{tid} and m_{tid} denote the number of individuals in demographic group d, county i and phase t who are in sample and who are overweight, respectively.

2.2. Estimation

We use a Bayesian approach to make inference about θ_{LD}. Bayesian inference is performed conditionally on the sampled individuals, their responses are known and need not be estimated. However, at our reference time of census day 1990, we do not know where any of our sampled individuals reside (except for the few that might have been interviewed on that day). Hence, we estimate overweight prevalence for the entire population (sampled and unsampled). Since our model includes a phase effect, we make predictions using $\Pr(Y_{idj} = 1 \mid p_{1id}, p_{2id}) = (p_{1id} + p_{2id})/2$ (that is, using proportion averaged over phase). After specifying a prior distribution for $\boldsymbol{\alpha}$ and Γ, we estimate the posterior mean and variance of θ_{LD} using MCMC methods.[3]

3. ESTIMATION METHODOLOGY WITH SAMPLE SELECTION ADJUSTMENT

Ignoring the way the sample was selected may produce erroneous inferences.[9] By including sufficient design information in a model, any design can become ignorable. Often, however, the design information needed for modelling and prediction is difficult or impossible to obtain. This is particularly true at the lower levels of sampling, where sampling frames are only constructed for the higher-level units in sample. Here, we do not attempt to incorporate within-PSU design characteristics in our model. Instead, for simplicity, we utilize non-ignorable design methodology.[9]

3.1. Sample Selection Model

Based on previous extensive data analysis, we conclude that the design is ignorable above the PSU-level so that the model, defined in (4) and (5), is not affected by the PSU-level design characteristics. Hence, we use non-ignorable design methodology within counties (PSUs), only.

Let π_{tidj} denote the selection probability of individual $tidj$, as specified in the sample design. The resulting 'empirical Bayes' likelihood L_{NI} based on a non-ignorable design is

$$L_{NI}(\{\boldsymbol{\beta}_i\}_{i\in s}, \boldsymbol{\alpha}, \Gamma) = \prod_{i\in s}\left[\prod_t \prod_d \frac{p_{tid}^{m_{tid}}(1-p_{tid})^{n_{tid}-m_{tid}}}{(p_{tid}/\bar{w}_{1td} + (1-p_{tid})/\bar{w}_{0td})^{n_{tid}}}\right] |\Gamma|^{-1/2} \mathrm{e}^{-1/2\boldsymbol{\beta}_i^{\mathrm{T}}\Gamma^{-1}\boldsymbol{\beta}_i} \tag{7}$$

where $\bar{w}_{1td}$ and $\bar{w}_{0td}$ are the sampling weights (inverse selection probabilities) for demographic group d in phase t, averaged over overweight and normal persons, respectively. That is

$$\bar{w}_{1td} = \sum_{(i,k)\in s_{td}} Y_{tidk}\pi_{tidk}^{-1} \Big/ \sum_{(i,k)\in s_{td}} Y_{tidk}$$

and

$$\bar{w}_{0td} = \sum_{(i,k)\in s_{td}} (1 - Y_{tidk})\pi_{tidk}^{-1} \Big/ \sum_{(i,k)\in s_{td}} (1 - Y_{tidk})$$

where $(i, k)\in s_{td}$ denotes all sampled persons in demographic group d in phase t. The sample adjusted likelihood in (7) differs from (6) only in the denominator which adjusts for oversampling of both overweight and normal persons. A brief description on how the denominator of (7) was derived is presented in the Appendix. Alternatively, the adjustment in the denominator of (7) can be viewed as conditioning on the sample selection (for example, see Jewell[10]).

A full Bayesian analysis includes a model for the distribution of the π_{tidj}'s as part of the likelihood, instead of substituting in their corresponding maximum likelihood estimates (MLEs), as we have done (in equation (7)). Our empirical Bayes analysis, although only an approximation to the full Bayesian analysis, was chosen for its relative simplicity. The likelihood in (7) when combined with an appropriate prior distribution enables one to make estimates for θ_{LD}.

4. INFERENTIAL METHODOLOGY

A Bayesian analysis requires the specification of a prior distribution for $(\boldsymbol{\alpha}, \Gamma)$. To ensure that the sample information dominates the inference, we used an overdispersed prior distribution. In particular, we choose the conditional density of $\boldsymbol{\alpha}\,|\,\Gamma$ to be constant and an inverse Wishart distribution for Γ with one degree of freedom and mean $= vI_{6\times 6}$. This prior (with $v = 10^4$) is dominated by the data but seems to avoid problems with the use of vague priors in hierarchical models.[11]

Since the posterior moments of θ_{LD} are non-linear functions, and the posterior distribution cannot be expressed in a simple form, numerical evaluation is needed. We used MCMC methodology to generate a random set of parameters which converges to a stationary distribution that is the posterior distribution. This is achieved by successively generating parameter subsets from their conditional distributions. We used graphics and formal statistical tests to determine when convergence to the stationary distribution was attained. The parameters after this point were treated as a sample from the posterior distribution and used to estimate posterior moments. More specifically, we used the block-at-a-time Metropolis–Hastings algorithm[12] to generate one long run of the chain. Since the explicit posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unknown, the modes and Hessians were searched at each iteration to determine the candidate-generating Gaussian densities. Conditionally Γ was sampled directly from its inverse Wishart distribution. We also used CODA software[13] to perform the output analysis and convergence diagnosis for the chain. Within CODA, we used the Heidelberger and Welch[14] test to determine the number of iterations to discard and to determine if the Markov process was indeed stationary.

5. ESTIMATION RESULTS

We illustrate the calculations using two choices of L and D from (2). In Section 5.1 we show the estimates made at the national level for demographic subgroups. We compare our model-based

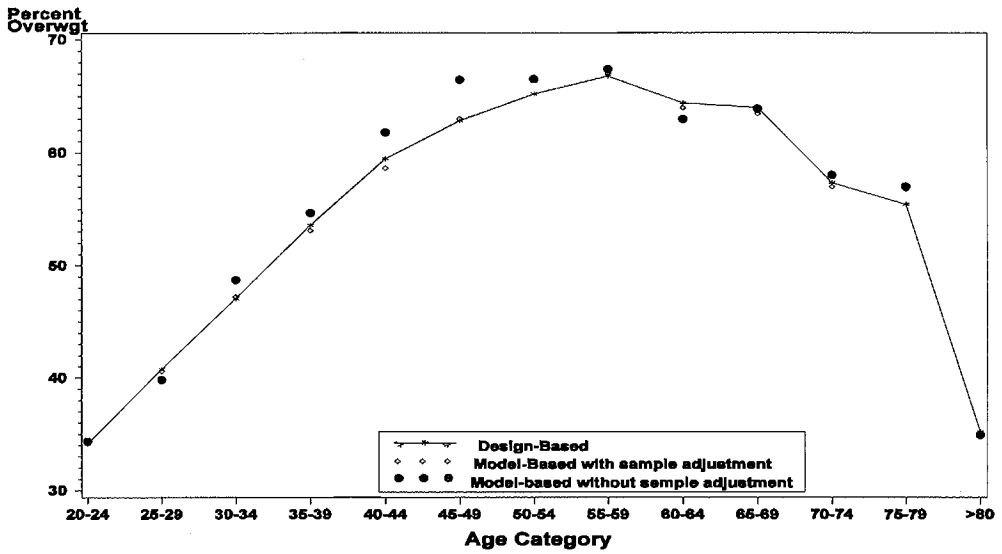


Figure 1. Comparison of estimates of overweight for non-Hispanic Black females at the national level: design-based versus model-based (with and without sample adjustment)

estimates with design-based estimates to justify our claim that the results may coincide at a national level. In Section 5.2 we provide comparison of our state level estimates with design-based and synthetic estimates. In Section 5.3 we show our estimates for all adults within the 50 states and D.C.

5.1. Evaluation of National Estimates by Demographic Subgroups

In this section we compare model- and design-based estimates for demographic categories. For estimation, we used the 16,523 BMI values for all adults (20 and over) who were examined in a mobile examination centre (MEC). We used standard expansion estimators to estimate the overweight prevalence for all demographic categories using the MEC examination weights.[15] Overweight prevalence is highest for ethnic (non-Hispanic Black and Mexican-American) females.[6]

For the model-based estimates, we approximated the selection probabilities of Section 3.1 by the inverse of the MEC examination weights after post-survey adjustment. We used SAS IML for the calculations and 1200 iterations of the Gibbs sampler. The estimates are based on the final 1000 iterations since the Heidelberger and Welch test indicates that the chain had converged by then. The values shown were obtained for the prior distribution with $v = 10^4$. We used sensitivity analysis to insure that our prior was overdispersed.

In Figure 1, we compare design-based, model-based without adjustment for sample selection, and model-based with adjustment for sample selection. Since the conclusions are similar for all race/ethnicity/gender categories, we illustrate our methodology using Non-Hispanic Black females. With adjustment for sample selection (specified in Section 3) the model-based estimates tracked the design-based estimates well for all ages. Without the sample selection adjustment, the

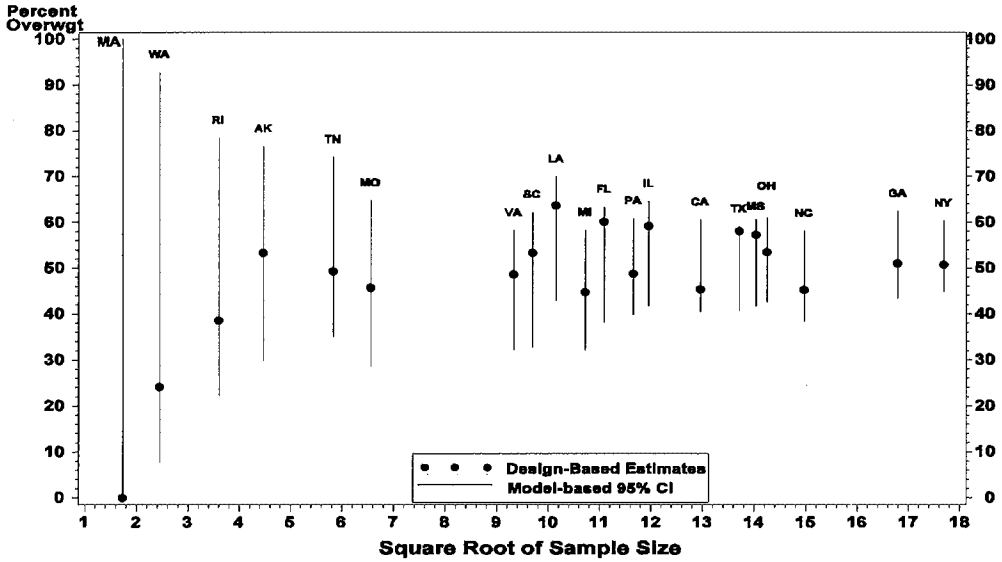


Figure 2. Comparison of state design-based estimates and model-based 95 per cent credible intervals (CIs) for overweight for non-Hispanic Black females

differences between the model-based and the design-based estimates were more pronounced (up to 5 percentage points difference). We take these results as verification that the sample selection adjustment was necessary.

5.2. Evaluation at the State Level

It is well known that hierarchical models, like the one used here, tend to 'smooth' estimates. Some smoothing is desirable because state estimates using only state data will contain more error and, hence, be more variable than estimates that 'borrow strength' from data outside of the state. However, there is concern that the model being used could inadvertently be smoothing the actual population values, not just removing error. If our model under represents the variability between counties, we will oversmooth. To check our model for oversmoothing, we evaluate the variability inherent in our model against the variability of the raw data as follows:

1. We first produce state design-based estimates for each state that contains at least one sampled PSU. These estimates use only data collected within the state and, hence, are not smoothed by averaging over data collected in other states.
2. Using our model with sample selection adjustment, we produce new values of overweight status for each sampled person via their posterior predictive distribution (with sample adjustment). We make corresponding state design-based estimates from these new outcomes. Since the new outcomes are from the posterior, estimates based on them will be smoothed.

Figure 2 compares the design-based state estimates with 95 per cent credible intervals[11] from our new outcomes for non-Hispanic Black females. Each raw design-based estimate falls within the credible interval formed using our model. If our model had oversmoothed, our generated

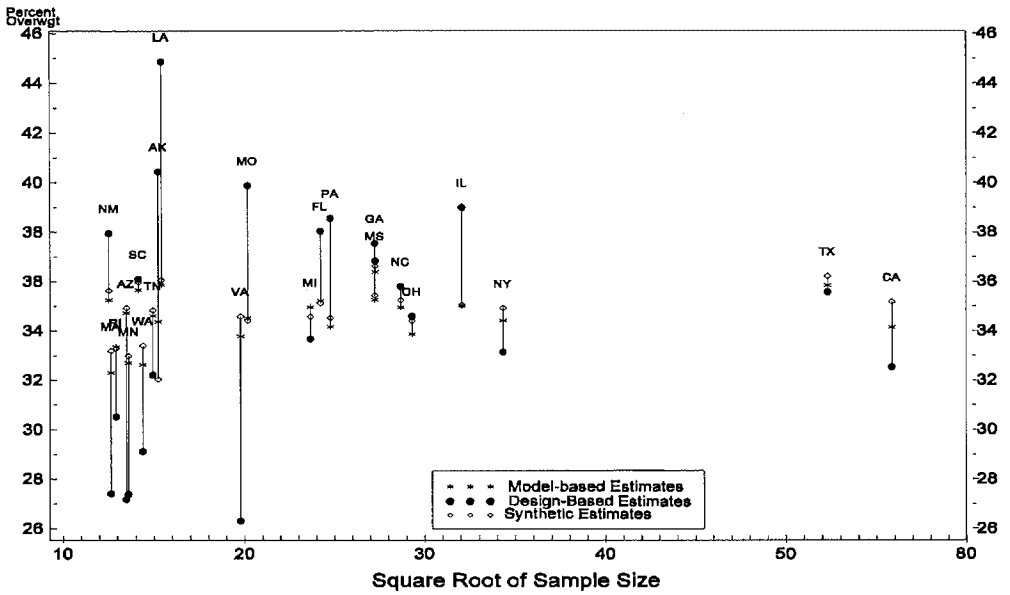


Figure 3. Comparison of model-based design-based and synthetic estimates of overweight for adults in sampled states

outcomes would tend to underestimate the data variability and would have resulted in poor coverage of the design-based estimates. We obtain this conclusion in the other race/gender categories, also.

Figure 3 compares the following three estimates for states that have a sample: model-based, design-based, and synthetic (defined by $\hat{\theta}_{LD} = \sum_{i\in L}\sum_{d\in D} N_{id}\hat{r}_d / \sum_{i\in L}\sum_{d\in D} N_{id}$, where $\hat{r}_d$ is the design-based estimate of the national prevalence rate for domain d).

The figure shows that, for states with a large sample, the model-based estimate is closer to the design-based estimate than is the synthetic estimate. In addition, the model-based posterior variance (not shown), decreases with state sample size. These two observations illustrate that the model-based approach preferentially uses state data. In general, the synthetic estimates are close to our estimates, suggesting that they are adequate in this case. However, without using the hierarchical model to account for between-county variation, we could not have made this conclusion. In addition, our method provides estimates of precision including county variation.

5.3. State Estimates

We computed the overweight prevalence estimate by state and show the results in Figure 4. The figure shows a relatively small range (0.32 to 0.40) and a north/south difference (reflecting the difference in minority population).

6. CONCLUSIONS

With the aim of producing accurate state estimates of overweight status, a hierarchical model with random county variation was employed. This model provides information on the

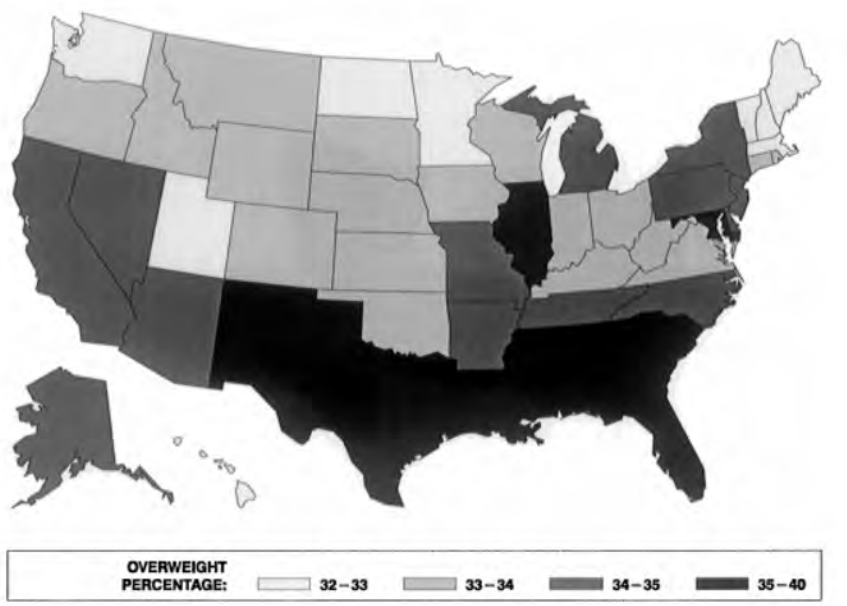


Figure 4. Overweight prevalence for adults by state

geographic variation of county prevalence rates and produces state estimates that preferentially use state data. The effects due to sample selection bias were minimized in three ways. First, county variables related to the NHANES III stratification variables were candidates in the stepwise variable selection procedure. For overweight status, however, these variables do not account for much of the overall data variability and were omitted from the model. Second, the county random effects component used to account for geographic variation also accounts for possible overdispersion due to the clustered sample. Third, the sample selection model was included in the likelihood to account for oversampling within PSUs.

Model misspecification was evaluated in two ways. First, we compared our estimates utilizing sample selection adjustment, with design-based estimates at the national level and obtained excellent agreement for all demographic groups. Second, we produced model-based credible intervals for the design-based state estimates and observed that these estimates fell within the intervals, indicating that the variability of these estimates had been adequately modelled. An unresolvable problem, however, in evaluating small area estimates is that accurate comparisons are not available at the small area level. In this application, for example, the design-based estimates at the state level are of poor precision and cannot be viewed as the 'gold standard'. Based on the model evaluations that we can make, we conclude that these estimates provide useful information on overweight prevalence for the states and D.C.

APPENDIX: SAMPLE SELECTION DERIVATION

We assume that the population of selection probabilities within a PSU, π_{tidj}, are identically distributed within demographic group, phase and overweight status. We specify an approximate

distribution for the selection probabilities by first assuming that they have support consisting only of the unique observed values $\pi_1^*, \ldots, \pi_U^*$. In particular, given that $Y_{tidj} = y$ $(y = 0, 1)$, π_{tidj} is modelled as a multinomial random variable with

$$\Pr(\pi_{tidj} = \pi_u^* \mid \theta, Y_{tidj} = y, p_{tid}) = \theta_{tdyu}. \tag{8}$$

Note that (8) makes few assumptions about the distribution of the selection probabilities. We have, however, assumed that the distribution of selection probabilities is identical between counties, within a demographic group and phase. We first show how the likelihood of overweight status is derived when $\boldsymbol{\theta}$ is fixed and then show how an estimate of $\boldsymbol{\theta}$ is derived to obtain (7).

We apply the general methodology for handling non-ignorable designs[9] to this case. Since our inference is conditional on county demographic groups and the NHANES III uses extensive implicit stratification, we assume each sampled person can be viewed as a sample of one person per substratum consisting of people in the same demographic group and phase. For sampled person j, we assume that the substratum size N_{tidj} is unknown. Dropping the extraneous subscripts (*tid*), for each person k in substratum j we let δ_{kj} be a Bernoulli random variable with sample selection probability π_{kj}. Without loss of generality, we label the sampled person as the first person in substratum j. The joint distribution of the one sampled person and the remaining unsampled persons in substratum j is

$$\begin{aligned}\Pr(\delta_{1j} = 1, Y_{1j}, \pi_{1j}, \{\delta_{kj} = 0, Y_{kj}, \pi_{kj}\}_{k=2,\ldots,N_j} | N_j) \\ = \pi_{1j}\Pr(\pi_{1j} | Y_{1j})\Pr(Y_{1j}) \prod_{k=2}^{N_j} (1 - \pi_{kj})\Pr(\pi_{kj} | Y_{kj})\Pr(Y_{kj}).\end{aligned} \tag{9}$$

The marginal distribution of only the observed quantities can be obtained by summing (9) over the unobserved components, giving

$$\begin{aligned}\Pr(\delta_{1j} = 1, Y_{1j}, \pi_{1j}, \{\delta_{kj} = 0\}_{k=2,\ldots,N_j} | N_j) \\ = \pi_{1j}\Pr(\pi_{1j} | Y_{1j})\Pr(Y_{1j}) \left(1 - \sum_{y=0}^{1} \sum_{u} \pi_u^* \Pr(\pi_u^* | y)\Pr(y)\right)^{N_j - 1}.\end{aligned} \tag{10}$$

Utilizing a non-informative prior, $\Pr(N_j) \propto$ constant, N_j can be removed from (10) giving

$$\begin{aligned}\Pr(\delta_{1j} = 1, Y_{1j}, \pi_{1j}, \{\delta_{kj} = 0\}_{k \neq 1}) &\propto \sum_{N_j=1}^{\infty} \pi_{1j}\Pr(\pi_{1j} | Y_{1j})\Pr(Y_{1j}) \left(1 - \sum_{y=0}^{1} \sum_{u} \pi_u^* \Pr(\pi_u^* | y)\Pr(y)\right)^{N_j - 1} \\ &\propto \pi_{1j}\Pr(\pi_{1j} | Y_{1j})\Pr(Y_{1j}) \Big/ \sum_{y=0}^{1} \sum_{u} \pi_u^* \Pr(\pi_u^* | y)\Pr(y).\end{aligned} \tag{11}$$

Equation (11) provides the likelihood component for one sampled person and replaces the specification in (3). Substituting the full subscript notation, for sample person j, into (11), one has

$$\begin{aligned}\Pr(\delta_{tidj} = 1, Y_{tidj} = y, \pi_{tidj} = \pi_u^*, \{\delta_{tidk} = 0\}_{k \neq j} | p_{tid}) \\ = \pi_u^* \theta_{tdyu} p_{tid}^y (1 - p_{tid})^{1-y} \Big/ \sum_{y=0}^{1} \sum_{u} \pi_u^* \theta_{tdyu} p_{tid}^y (1 - p_{tid})^{1-y} \\ \propto p_{tid}^y (1 - p_{tid})^{1-y} \Big/ \sum_{y=0}^{1} \sum_{u} \pi_u^* \theta_{tdyu} p_{tid}^y (1 - p_{tid})^{1-y}.\end{aligned} \tag{12}$$

By including the distribution of the selection probabilities specified in (8) as part of the likelihood, we could perform a complete Bayesian analysis. For simplicity, we substitute the MLE of $\boldsymbol{\theta}$ (based on conditioning on overweight status and the selection probability) into (12). For sampled individual *tidj*, the component of the conditional likelihood is

$$\Pr(\pi_{tidj} = \pi_u^* \mid \delta_{tidj} = 1, Y_{tidj} = y, \boldsymbol{\theta}) = \frac{\Pr(\delta_{tidj} = 1 \mid \pi_{tidj} = \pi_u^*, Y_{tidj} = y, \boldsymbol{\theta}) \Pr(\pi_{tidj} = \pi_u^* \mid Y_{tidj} = y, \boldsymbol{\theta})}{\Pr(\delta_{tidj} = 1 \mid Y_{tidj} = y, \boldsymbol{\theta})}$$

$$\propto \frac{\theta_{tdyu}}{\sum_u \pi_u^* \theta_{tdyu}}. \tag{13}$$

Multiplying all the terms of (13) together, the conditional likelihood is $\prod_{tdyu} (\theta_{tdyu} / \sum_u \pi_u^* \theta_{tdyu})^{\tau_{tdyu}}$ where τ_{tdyu} denotes the sample frequency of π_u^* within demographic group d, phase t and overweight status y. It can be shown that the MLEs of the θ_{tdyu} are

$$\hat{\theta}_{tdyu} = \frac{\tau_{tdyu}/\pi_u^*}{\sum_u \tau_{tdyu}/\pi_u^*}. \tag{14}$$

Plugging the estimates from (14) into (12) completes the within-PSU likelihood component in (7).

ACKNOWLEDGEMENTS

We want to thank Kurt Maurer for helpful comments and interest in the NHANES III small areas estimation project, Joe Sedransk for providing helpful comments on an earlier draft, and the referees and editor for providing comments that greatly improved the presentation of results.

REFERENCES

1. Ghosh, M. and Rao, J. N. K. 'Small area estimation: an appraisal', *Statistical Science*, **9**, 55–93 (1994).
2. Malec, D., Sedransk, J., Moriarity, C. and LeClere, F. 'Small area inference for binary variables in the National Health Interview Survey', *Journal of the American Statistical Association*, **92**, 815–826 (1997).
3. Gilks, W. R., Richardson, S. and Spielgelhalter, D. (eds). *Practical Markov Chain Monte Carlo*, Chapman and Hall, New York, 1996.
4. National Center for Health Statistics. Plan and operation of the Third National Health and Nutrition Examination Survey, 1988–94, *Vital Health Statistics*, **1**, (1994).
5. Troiano, R. P., Frongillo, E. A. Jr., Sobal, J. and Levitsky, D. A. 'The relationship between body weight and mortality: a quantitative analysis combining information from existing studies', *International Journal of Obesity*, **20**, 63–75 (1996).
6. Kuczmarski, R. J., Flegal, K. M., Campbell, S. M. and Johnson, C. L. 'Increasing prevalence of overweight among US adults: the National Health and Examination Surveys, 1960 to 1991', *Journal of the American Medical Association*, **272**, 205–211 (1994).
7. Galuska, D. A., Serdula, M., Panuk, E., Siegel, P. Z. and Byers, T. 'Trends in overweight among US Adults from 1987 to 1993: a multistate telephone survey', *American Journal of Public Health*, **86**, 1729–1735 (1996).
8. Scott, A. J. 'On the problem of randomization in survey sampling', *Sankhya*, **39**, Series C, Pt.1, 1–9 (1977).
9. Gelman, A., Carlin, B. B., Stern, H. S. and Rubin, D. B. *Bayesian Data Analysis*, Chapman and Hall, New York, 1995, Chapter 7.
10. Jewell, N. P. 'Least squares regression with data arising from stratified samples of the dependent variable', *Biometrika*, **72**, 11–22 (1985).
11. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn, Springer-Verlag, New York, 1986, Chapter 4.

12. Chib, S. and Greenberg, E. 'Understanding the Metropolis-Hastings algorithm', *American Statistician*, **49**, 327–335 (1995).
13. Best, N., Cowles, M. K. and Vines, K. *CODA*: *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*, Version 0.30, MRC Biostatistics Unit, Cambridge, 1995.
14. Heidelberger, P. and Welch, P., 'Simulation run length control in the presence of an initial transient', *Operations Research*, **31**, 1109–1144 (1983).
15. Mohadjer, L., Montaquila, J., Waksberg, J., Bell, B., James, P., Flores-Cervantes, I. and Montes, M. 'National Health and Nutrition Examination Survey III: Weighting and Estimation Methodology', Westat Inc., Rockville, MD, 1996.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3201–3209 (1999)

APPLICATION OF A WEIGHTED HEAD-BANGING ALGORITHM TO MORTALITY DATA MAPS†

MICHAEL MUNGIOLE[1]*, LINDA W. PICKLE[1] AND KATHERINE HANSEN SIMONSON[2]

[1] *Centers for Disease Control and Prevention, National Center for Health Statistics, 6525 Belcrest Rd., Rm. 915, Hyattsville, MD 20782-2003, U.S.A.*
[2] *Sandia National Laboratories, PO Box 5800, Albuquerque, NM 87185-0844, U.S.A.*

SUMMARY

Smoothed data maps permit the reader to identify general spatial trends by removing the background noise of random variability often present in raw data. To smooth mortality data from 798 small areas comprising the contiguous United States, we extended the head-banging algorithm to allow for differential weighting of the values to be smoothed. Actual and simulated data sets were used to determine how head-banging smoothed spike and edge features in the data, and to observe the degree to which weighting affected the results. As expected, spikes were generally removed while edges and clusters of high rates near the U.S. borders were maintained by the unweighted head-banging algorithm. Incorporating weights inversely proportional to standard errors had a substantial effect on smoothed data, for example determining whether observed spikes were retained or removed. The process used to obtain the smoothed data, including the choice of head-banging parameters, is discussed. Results are considered in the context of general spatial trends. Published in 1999 by John Wiley & Sons, Ltd. This article is a U.S. Government work and is in the public domain in the United States.

1. INTRODUCTION

When looking at mortality data maps, the reader is often unable to discern any spatial trends, either because no particular patterns exist in the data, or because a substantial amount of noise (that is, spatial variability) masks the patterns that are present. For example, Lewandowsky and Behrens[1] found that the consistency of detection of high rate clusters on a simulated map was hampered as an increasing amount of noise was added to the data. Statistical smoothing algorithms can be used to enhance hidden patterns in noisy data. This is accomplished by removing local small-scale variations while maintaining larger-scale regional trends.

Two types of features that may influence one's ability to recognize broad trends in spatial data are spikes and edges. A spike represents an isolated extreme value (peak or depression) that contrasts greatly with other nearby data points. Spikes may be caused by measurement error and represent inherent noise in the mapped variable or they may be legitimate values, for example, a high rate for a city that is surrounded by suburbs with lower rates. Edges are zones of rapid transition in the variable of interest. They may appear as ramps, valleys, or ridges and generally represent legitimate structure in the data. In fact, edges can be particularly valuable in helping the

* Correspondence to: Michael Mungiole, Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783-1197, U.S.A. E-mail: mmungiole@mail.arl.mil
† This article is a U.S. Government work and is in the public domain in the United States.

Published in 1999 by John Wiley & Sons, Ltd.

reader to identify spatial regions that are distinct from one another in terms of the mapped variable.

In applications such as the various geologic examples discussed in Hansen[2], spikes are physically unlikely and it is thus desirable for smoothers to remove all spikes while retaining the edges. When working with mapped statistical data, however, we can often distinguish between spikes that are credible (for example, rates in cities with large populations) and those that are of suspect origin (likely due to random noise). We would like our smoother to retain the former (and the edges as well) while removing the latter.

Many authors have observed that the use of mean-based smoothers results in the blurring of both the spikes and edges for one-dimensional time series data[3] and for spatial data.[2,4,5] Median-based approaches often represent a reasonable alternative, removing spikes while preserving the sharpness of edges. Tukey[6] and Velleman[3] propose a variety of median-based smoothers for time series data, while Jain[4] discusses non-linear filters for two-dimensional lattices.

The problem of smoothing non-gridded spatial data has received attention from a variety of authors. Kafadar[5] discusses a number of the techniques that have been proposed, and evaluates their performance on several constructed data sets using an overall mean squared error statistic. The results of this comparative study suggest that the weighted average (based on inverse squared distance) and the median-based head-banging smoother, proposed by Tukey and Tukey,[7] and implemented as in Hansen,[2] perform well. However, the weighted average and most other smoothers included in the comparison tended to oversmooth in the presence of edges, whereas head-banging was able to preserve their steep structure.

To utilize the best features of these smoothers, we decided to use a modified head-banging approach in smoothing mortality data for an atlas of maps of numerous small areas in the contiguous United States.[8] Because the original implementation of head-banging[2,7] does not allow for the use of reliability estimates during the smoothing process, we have developed an enhanced version, in which the user may provide a set of weights along with the mapped data. We refer to this new technique as *weighted head-banging* (a copy of the weighted head-banging code, written in the C language, is available from K. Simonson), and demonstrate its application to mortality data.

2. METHODS

2.1. Notation: Weighted Medians

Suppose that a mapped data set contains a total of n data points. Each point, i, consists of a location in the plane, $\mathbf{x}_i = (x_{i1}, x_{i2})$, which is assumed to be known without error, along with a (possibly noisy) observation, y_i, and a weight, w_i, which measures the reliability of y_i. The goal is to smooth the mapped values $\{y_i, i = 1, \ldots, n\}$, accounting for both location and reliability. Our approach assigns a smoothed value at location $\mathbf{x}_i$ that is based on weighted medians of the observations at neighbouring locations and the observation y_i itself.

Given m values $\{y_1, y_2, \ldots, y_m\}$, with corresponding weights $\{w_1, w_2, \ldots, w_m\}$, the *weighted median* of the y's is found as follows. First, sort the observations, so that $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(m)}$. The weights, which remain with their original observations, are re-ordered such that $w_{(j)}$ corresponds to the sorted observation $y_{(j)}$. Now compute the cumulative sums of the re-ordered weights:

$$S_k = \sum_{j=1}^{k} w_{(j)}, \quad k = 1, \ldots, m. \tag{1}$$

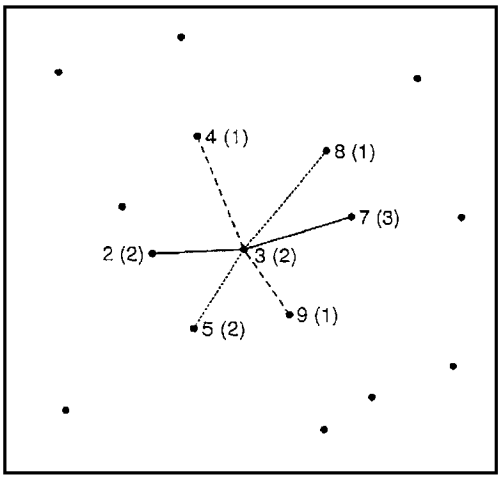


Figure 1. Three nearly collinear triples are used for smoothing at this centre point. Observed values are printed at each data location, with corresponding weights given in parentheses. The low screen is *weighted median* (2, 4, 5) = 4. The high screen is *weighted median* (7, 8, 9) = 7. The observed value at the centre point lies below the low screen, so the weights are used to determine whether adjustment is needed. Because the sum of endpoint weights (10) exceeds the number of triples times the centre point weight ($3 \times 2 = 6$), the smoothed value at the centre point is set to the low screen

The index location, r, of the weighted median is the smallest value of k such that S_k is at least half of the total sum of the weights:

$$r = \min \{k | S_k \geqslant S_m/2\}. \tag{2}$$

If S_r is strictly greater than $S_m/2$, then the weighted median value is given by $y_{(r)}$. If S_r is equal to $S_m/2$, then the weighted median value is taken to be the average of $y_{(r)}$ and $y_{(r+1)}$.

2.2. The Weighted Head-Banging Smoother

In one-dimensional running median smoothers, the output value at a given point is the median of the observed value at that point, and a specified number of neighbouring points on both the left and right sides. Head-banging attempts to bring the simplicity of this approach to spatial data, where left and right neighbours are not well defined. Triples of nearby data points are used in their place.

For each point in the data set, a collection of nearly collinear triples is defined. Every triple in the collection consists of three data points in the same region of the plane, and is centred at the location of the point to be smoothed. A method for identifying such collections is discussed in the next section; for now assume that at least one triple has been chosen for each point in the data set. The iterative weighted head-banging smoother is carried out as illustrated in Figure 1.

For each i, set the initial estimate for location $\mathbf{x}_i$ at its observed value. Let T_i represent the collection of triples centred at $\mathbf{x}_i$. Now for the jth triple in T_i, let low_{ij} represent the smaller of the two endpoint values, and let high_{ij} represent the larger. Define the following two quantities:

$$\text{high screen } (i) = \text{weighted median}_j \, (\text{high}_{ij})$$

$$\text{low screen } (i) = \text{weighted median}_j \, (\text{low}_{ij}).$$

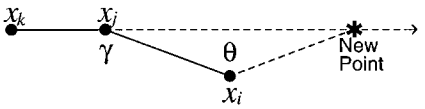


Figure 2. Extrapolating triples for an edge or corner point. If no qualifying triples are centred at point $\mathbf{x}_i$, new triples are formed by extrapolating the y-values from two neighbouring locations, $\mathbf{x}_j$ and $\mathbf{x}_k$. These values are linearly extrapolated to a point lying along the line determined by $\mathbf{x}_j$ and $\mathbf{x}_k$, such that $\mathbf{x}_i$ is equidistant from $\mathbf{x}_j$ and the new point [after Hansen[2]]

The low screen is an estimate of the lowest value that is consistent with the data in the neighbourhood of the centre point, and the high screen is an estimate of the highest such value.

If the current centre value lies between these screens, it is left unchanged at this iteration. If the current centre value is smaller than the low screen, the weights are used to determine whether it will be adjusted. Specifically, if the sum of the weights of all endpoints of the triples exceeds the number of triples times the weight of the centre point, then the smoothed value at the centre point takes the value of the low screen. Similarly, if the current centre value is larger than the high screen, and the summed endpoint weights exceed the number of triples times the centre weight, the smoothed centre value is set equal to the high screen. Even if the current centre value is changed to the low or high screen, it still retains its original weight.

This process is carried out for each point in the data set and all data points are updated simultaneously at the end of each iteration. Thus, the order in which points are smoothed does not influence the values resulting from a particular iteration. The procedure is repeated for a fixed number of iterations, or until no further changes take place. When equal weights are assigned to each data point, this procedure is identical to that outlined in Tukey and Tukey.[7]

2.3. Selection of Triples

Two general criteria for selecting triples are as follows: the three points should be located close to one another, so that smoothing takes place on a local or regional rather than a global scale; and the three points should be roughly collinear, so that directional trends can be recognized and enhanced. An algorithm for identifying collections of triples meeting these criteria is described in Hansen[2] and has been adopted here.

The regional criterion is satisfied by requiring that each of the endpoints of a candidate triple be among the NN nearest neighbours of the centre point. The scalar NN is a parameter chosen by the user; increasing NN will increase the degree of smoothness in the final map. The directional criterion is specified in terms of θ, the centre angle. The user selects a threshold, θ^*, and those triples for which $\theta \geqslant \theta^*$ are considered candidates. We have always set $\theta^* = 135^\circ$; this gives good results over a variety of applications.

For many centre points, the collection of triples eligible under these two criteria is sizable, and can be reduced by placing a limit, NTRIP, on the total number of selected triples. From the eligible set, the NTRIP thinnest (defined in terms of the perpendicular distance from the centre point to the line passing through the two endpoints) are retained. The use of NTRIP imposes an additional collinearity constraint on triples, which becomes more restrictive as the endpoints move away from the centre point. It also reduces the run-time of the full smoothing procedure.

The method discussed above may not identify any eligible triples for data points located at the edges and corners (perimeter) of a map. For such points, triples can be formed by extrapolation from pairs of neighbouring points, as illustrated in Figure 2. Suppose that the observation at $\mathbf{x}_i$ is

to be smoothed, and that locations $\mathbf{x}_j$ and $\mathbf{x}_k$ (with observed values y_j and y_k, respectively) are among its NN nearest neighbours. Consider the triangle with vertices at $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$, and let γ be the angle at $\mathbf{x}_j$. If $\gamma \geqslant 90^\circ + \theta^*/2$, then a new triple is formed by linearly extrapolating the trend from y_k through y_j to a new location such that $\mathbf{x}_i$ is equidistant from the new location and $\mathbf{x}_j$. This new location is one endpoint of the created triple, and $\mathbf{x}_j$ is the other. The restriction on γ ensures that the created triple has a centre angle that is greater than or equal to θ^*.

The weight assigned to the extrapolated point is equal to the smaller of the two weights, w_j and w_k. Once a collection of extrapolated triples has been identified for each edge and corner location, smoothing proceeds in the same manner as previously described.

2.4. Data Sets

The data used to illustrate the technique are mortality for lung and prostate cancer for 1988–1992 among White males from the recently-published NCHS *Atlas of United States Mortality*.[8] These causes were chosen because lung cancer has a pronounced ridge whereas prostate cancer rates exhibited a high degree of scatter across the small areas. These data consisted of directly age-adjusted death rates[9] and their variances for 798 geographic units in the contiguous U.S. The geographic unit that was mapped, health service areas (HSA), represents aggregates of counties based on where residents obtain their routine hospital care.[10] All data were mapped as classed choropleth (area shaded) maps using a monochromatic colour scheme; rates were classified into five categories according to quintiles of the presmoothed distribution. We considered the inverse of standard error of the rates as weights when smoothing the mortality data. In addition, simulated data which included edge and spike effects were provided by K. Kafadar for comparison.[5]

3. RESULTS

3.1. Lung Cancer Mortality Data

Plate 1 illustrates how weighted and unweighted head-banging differ in their smoothing characteristics. Each of the maps shown contain the same ranges (cutpoints) for all five classes to allow valid map comparisons. Plate 1(*a*) represents the original lung cancer mortality rate data (prior to smoothing); Plates 1(*b*) and 1(*c*) show the results of smoothing with the unweighted and weighted algorithms, respectively, using parameter values of NN = 12 and NTRIP = 8. An important requirement in smoothing mortality maps, as previously mentioned, is retaining the edges that represent valid data. The pronounced ridge running along the Ohio valley in a northeast-to-southwest direction on the original map (Plate 1(*a*)), is maintained as the map data is smoothed using both the weighted and unweighted head-banging algorithms.

Another feature of both algorithms is a narrowing of the distribution of mortality rates, due to the attenuation of outliers which are typically based on sparse data. Evidence of this distributional change in the lung cancer maps is the smaller range of rates seen for the smoothed maps as compared to the raw data map. Because none of the lung cancer outlier rates is based on large populations, the resulting ranges from both smoothing algorithms are very similar.

While it may not be apparent when viewing a small map of the contiguous U.S., one region that exhibits a difference between weighted and unweighted smoothing is north central Tennessee (Plate 2), which includes the Nashville metropolitan area. This densely-populated region has a small standard error and, hence, a large weight. Without weighting, the Nashville rate is

smoothed from the second into the highest (darkest) class because of the influence of high rates in the surrounding areas. With weighting, the Nashville rates remain closer to the original data values because of their relatively high weights as compared to surrounding areas.

A comparison of the perimeter of the U.S. on the original map (Plate 1(*a*)) with either of the smoothed maps (Plates 1(*b*) or 1(*c*)) indicates the moderate changes in the values for each HSA after being smoothed. Thus, unlike many other smoothers that may cause undesirable effects on the edges, head-banging results in areas near the edges retaining their original values to a large extent.

To compare the effects of the degree of smoothing, we applied the unweighted algorithm to the lung cancer mortality data (Plate 3), varying the number of nearest neighbours (NN) and the maximum number of triples (NTRIP) while the minimum triple angle (135°) and the number of iterations (10) were kept constant. Use of the unweighted algorithm allows one to observe the effects of smoothing parameters without the potential confounding effects of different weights. Although there are not major differences between these two smoothed maps, closer inspection in the southeastern U.S. indicates at least three areas where distinctions do exist: east-central Texas, the eastern border between Florida and Georgia, and part of the border between North and South Carolina. These areas remain in the highest class (darkest saturation) when a small number of nearest neighbours (8) and maximum triples (5) are used but move into the second highest class when NN and NTRIP are increased to values of 16 and 11, respectively. Because a majority of the additional HSAs used for this greater degree of smoothing also happen to be in the second highest class, they reduce the three indicated areas into this class.

3.2. Prostate Cancer Mortality Data

In an effort to determine how weighted head-banging smooths data containing a significant amount of noise, we applied the weighted algorithm to prostate cancer data (Plate 4). The raw data indicates the large number of spikes (peaks), which are primarily located in the central part of the U.S. These isolated high rates (spikes) in rural HSAs that are surrounded by lower rate HSAs are often smoothed out when weighted head-banging is applied. This is a good example of the head-banging algorithm drawing a more apparent spatial pattern from scattered raw data.

3.3. Simulated Structure Data

We also examined simulated data to extend the results of Kafadar[5] using the weighted algorithm. The height of a predetermined structure (Figure 3), representing the measured value, y, was generated and shown on a three-dimensional perspective graph. This structure includes a ridge, peak, and depression in known locations of an area representing 86 western U.S. counties. Gaussian error of 10 per cent (normalized by the weight for each respective county) was added to the structure (Plate 5(*a*)) and the result was smoothed using parameter values of 10 and 7 for NN and NTRIP, respectively, for the unweighted (Plate 5(*b*)) and weighted (Plate 5(*c*)) head-banging algorithms. Because standard errors were not available, the weights (Plate 5(*d*)) were approximated by the square root of county population. The unweighted algorithm (Plate 5(*b*)) retains the height of the region towards the back of the ridge, a 'compromise' between values at the front of the ridge and values on either side of the back of the ridge. However, weighting reduces the height of this low-weighted region to a level similar to nearby low-weighted areas (Plate 5(*c*)) because the low weights (Plate 5(*d*)) are nearly equal for counties along and on either side of the back portion of the ridge (Figure 3). Another difference between the weighted and unweighted head-banging

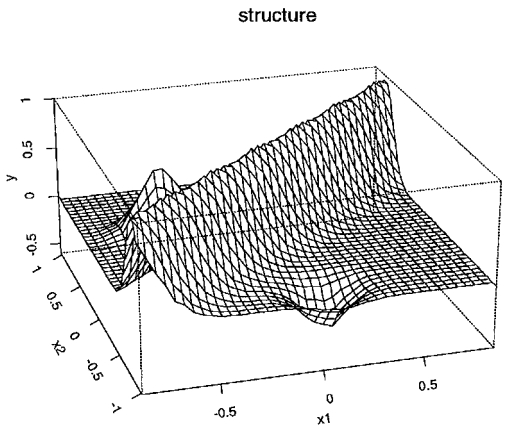


Figure 3. Simulated structure containing a ridge, peak and depression

Table I. Mean squared errors (and SE) between structure and structure plus noise (unsmoothed) and between structure and smoothed structure plus noise for unweighted and weighted head-banging. Results are given at three different noise levels

Condition	Noise level		
	10%	25%	50%
Unsmoothed	12.13 (0.412)	75.80 (2.573)	303.2 (10.29)
Unweighted head-banging	2.994 (0.0536)	6.687 (0.178)	17.66 (0.715)
Weighted head-banging	2.931 (0.0551)	6.129 (0.159)	13.74 (0.584)

results is that the peak is retained to a greater degree (that is, less height reduction) for the weighted algorithm due to the influence of the weights in this region.

The perspective graphs (Plates 5(*a*), (*b*), (*c*) and (*d*)) are also shown as choropleth maps (Plates 5(*e*), (*f*), (*g*) and (*h*), respectively), with the same orientation as the graphs. For comparison, the three maps of structure height have the same colours and cutpoint values for each of the five classes. Categorization of the structure height into five classes masks differences between the unweighted and weighted smoothed maps that are apparent in perspective graphs.

The performance of the unweighted and weighted head-banging smoothing of the structure plus noise were compared by calculating a mean squared error (MSE) between the original structure and the smoothed results, similar to the methods employed by Kafadar.[5] These results, for three different levels of noise, are given in Table I. For comparison, the MSE between the structure and structure plus noise are also included in this table. The weighted algorithm captures

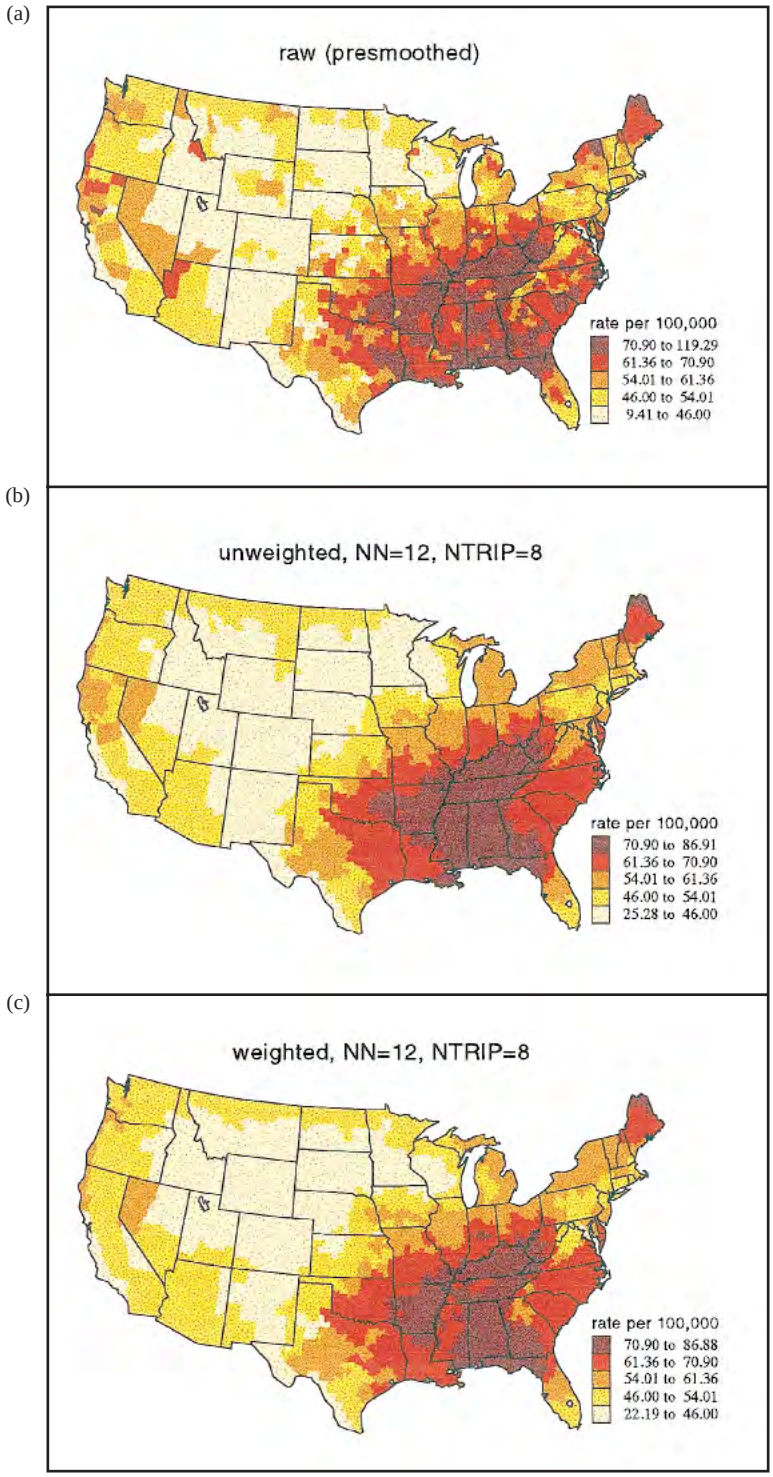


Plate 1. Comparison of weighted and unweighted algorithms for lung cancer mortality data

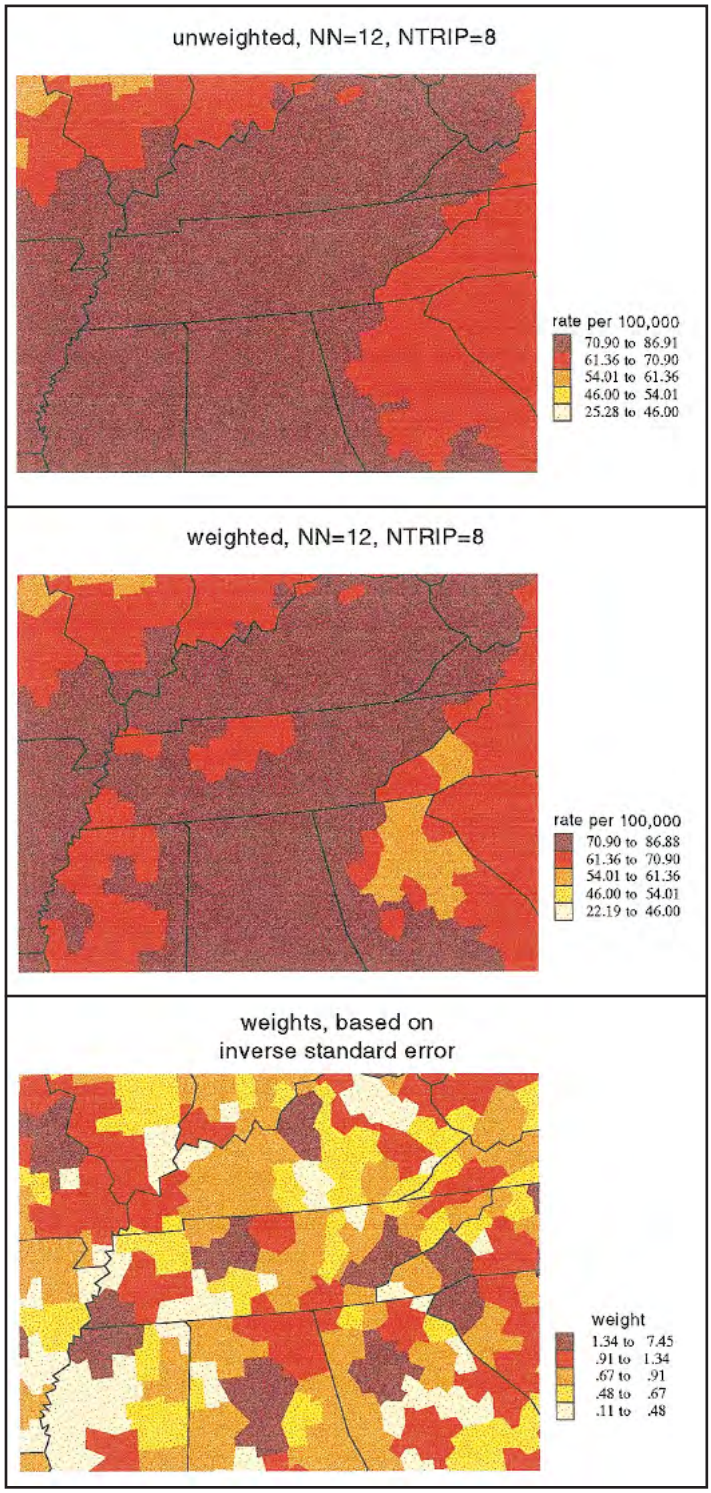


Plate 2. Comparison of the degree of smoothing between weighted and unweighted algorithms for a highly populated area in the vicinity of less populated areas

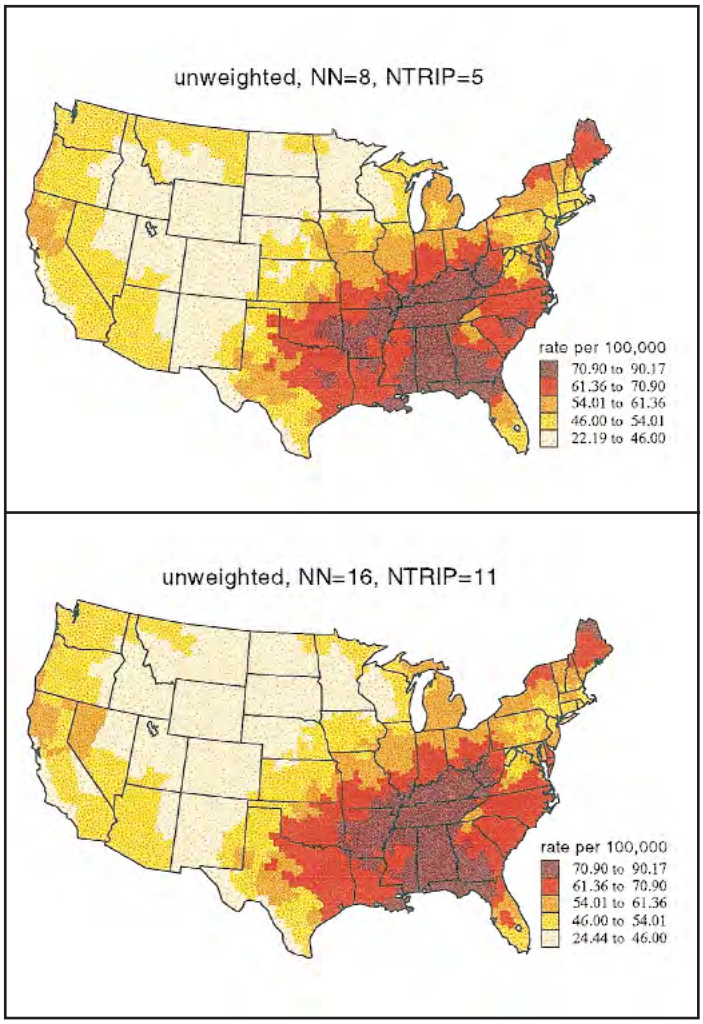


Plate 3. Influence of smoothing parameters for lung cancer mortality data, using unweighted head-banging

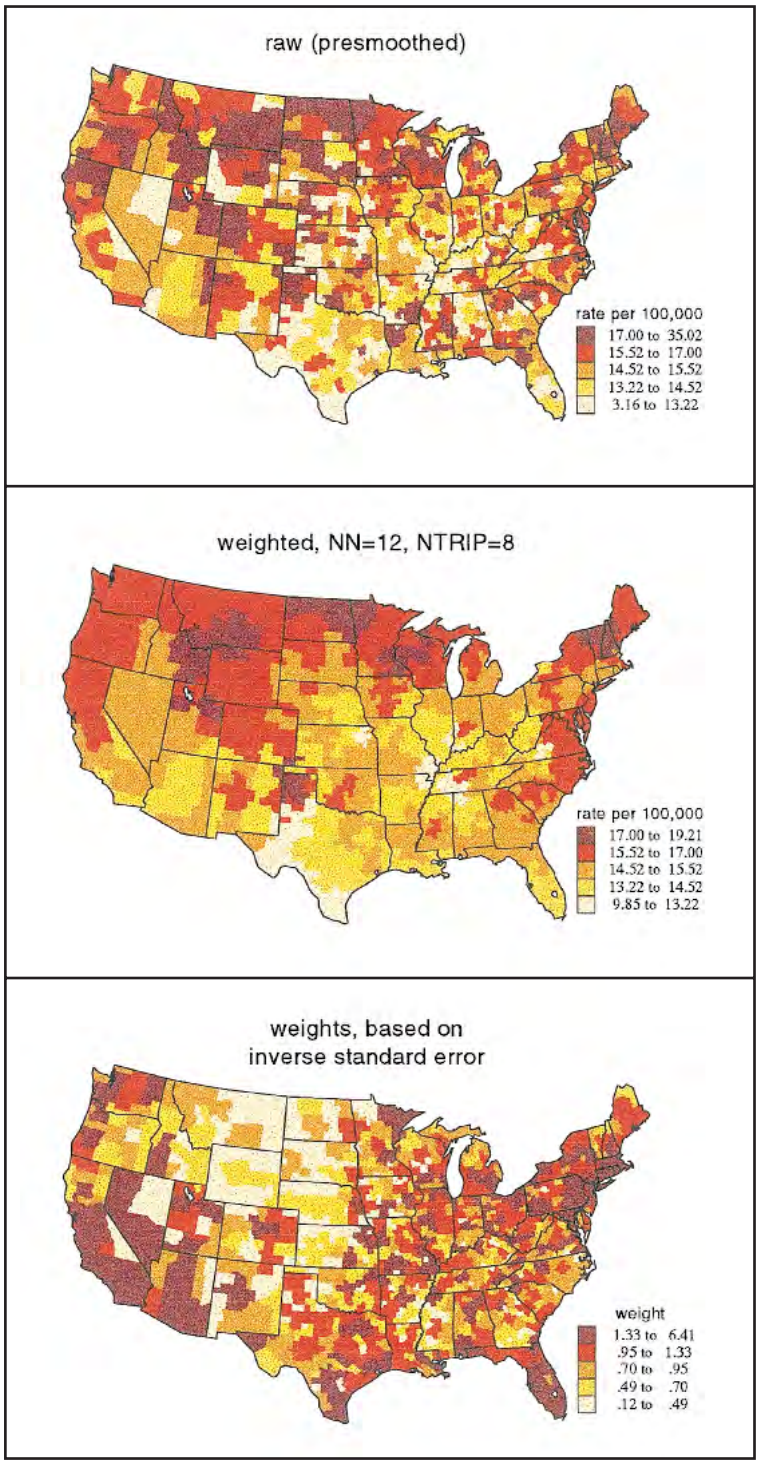


Plate 4. Effect of the weighted algorithm on 'noisy' prostate cancer mortality data

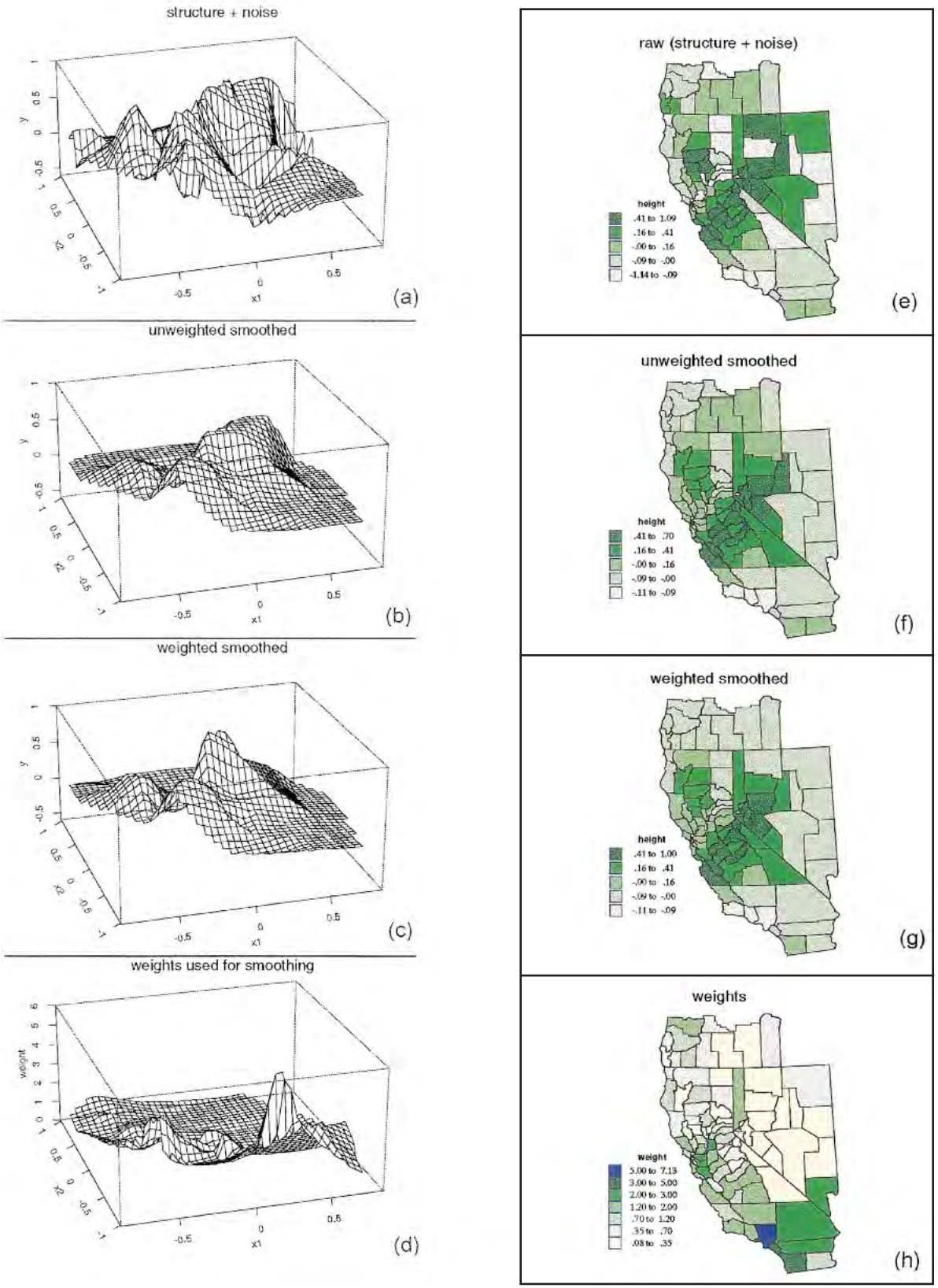


Plate 5. Effect of weighted and unweighted head-banging algorithms when used to smooth a simulated structure with added noise

the original structure better for the highest level of noise, while there is little difference between weighted and unweighted head-banging for the two lower levels of noise.

4. DISCUSSION

These results demonstrate the utility of using the head-banging smoothing algorithm for mortality rate maps, but the conclusions should apply to any choropleth maps of statistical data which are of varying reliability. Application of this algorithm to maps of lung cancer mortality and simulated data suggests that this technique retains edge effects both within the map and along its perimeter. Furthermore, the addition of weighting to the algorithm allows the degree of smoothing to depend on the variability of the mapped statistic. This feature produces greater smoothing of data that are less reliable, while retaining nearly original values of statistics based on large sample sizes.

For the prostate cancer mortality data, the head-banging smoother highlighted existing geographic patterns that were somewhat masked due to the high degree of variability of the rates. Both the weighted and unweighted algorithms removed the spikes of the many high rate HSAs in less populous areas. In contrast, high rates of HIV mortality nearly always occur in urban areas, where the population is large. Unpublished results indicate that these values are maintained with the weighted algorithm while they are often smoothed away for the unweighted case.

The weighted head-banging algorithm was better at recovering the structure from simulated data than the unweighted procedure, due to the reduced influence of the low-weighted counties when smoothing using the weighted algorithm. The improvement of the weighted algorithm ranged from a 22.2 per cent reduction in the MSE with 50 per cent noise, to slightly greater than a 2 per cent reduction for the 10 per cent noise level.

Selection of the appropriate parameters for smoothing these mortality and simulated data was subjective. It should be noted that the optimal parameters for other data sets may deviate from the values we selected and would be based on such factors as the number of data values to be smoothed, the noisiness in the data, and the distribution of the weights. Users should experiment with a number of parameter values to determine the effect of the weighted head-banging smoother on their data. Further research needs to be conducted to obtain general guidelines for choosing optimal parameter values.

In summary, results of applying the weighted head-banging algorithm to both actual and simulated statistical data for small geographic areas indicate its superiority for revealing underlying spatial patterns in the data. Unlike many other proposed smoothers, important features such as edges and spikes based on adequate data are retained, but unreliable data points are smoothed toward adjacent, more reliable values. The addition of weights extends the head-banging algorithm beyond applications where homogeneity of variance may be assumed, such as for maps of death rates. The modified algorithm should prove to be a useful tool for examining mapped data where statistical variability may mask important spatial structure.

REFERENCES

1. Lewandowsky, S. and Behrens, J. T. 'Accuracy of cluster detection in mortality maps', in Pickle, L. W. and Herrmann, D. J. (eds), *Cognitive Aspects of Statistical Mapping*, NCHS Working Paper Series, No. 18, National Center for Health Statistics, Hyattsville, MD, 1995.
2. Hansen, K. M. 'Head-banging: robust smoothing in the plane', *IEEE Transactions on Geoscience and Remote Sensing*, **29**, 369–378 (1991).

3. Velleman, P. F. 'Definition and comparison of robust nonlinear data smoothing algorithms', *Journal of the American Statistical Association*, **75**, 609–615 (1980).
4. Jain, A. K. *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.
5. Kafadar, K. 'Choosing among two-dimensional smoothers in practice', *Computational Statistics and Data Analysis*, **18**, 419–439 (1994).
6. Tukey, J. W. *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.
7. Tukey, P. A. and Tukey, J. W. 'Graphical display of data sets in 3 or more dimensions', Barnett, V. (ed.), *Interpreting Multivariate Data*, Wiley, New York, 1981.
8. Pickle, L. W., Mungiole, M., Jones, G. K. and White, A. A. *Atlas of United States Mortality*, National Center for Health Statistics, Hyattsville, Maryland, 1996.
9. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, Wiley, New York, 1981.
10. Makuc, D. M., Haglund, B., Ingram, D. D. *et al. Health Service Areas for the United States*, Vital Health Statistics 2(112), National Center for Health Statistics, Hyattsville, MD, 1991.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3211–3220 (1999)

EXPLORING SPATIAL PATTERNS OF MORTALITY: THE NEW *Atlas of United States Mortality*†

LINDA W. PICKLE[1]*, MICHAEL MUNGIOLE[2], GRETCHEN K. JONES[2] AND ANDREW A. WHITE[3]

[1] *NCI/DCCPS, 6130 Executive Boulevard, MSC 7344, EPN Rm. 313, Bethesda, MD 20892, U.S.A.*
[2] *National Center for Health Statistics, 6525 Belcrest Rd., Rm. 915, Hyattsville, MD 20782, U.S.A.*
[3] *National Academy of Sciences, 2101 Constitution Ave., N.W., Washington, D.C. 20418, U.S.A.*

SUMMARY

The National Center for Health Statistics, CDC, has produced an *Atlas of United States Mortality* which includes maps of rates for the leading causes of death in the United States for the period 1988–1992. As part of this project, many aspects of statistical mapping have been re-examined to maximize the atlas's effectiveness in conveying accurate mortality patterns to epidemiologists and public health practitioners. Because recent cognitive research demonstrated that no one map style is optimal for answering many different map questions, maps and graphs of several different mortality statistics are included for each cause of death. New mixed effects models were developed to provide predicted rates and improved variance estimates. Results from these models were smoothed using a weighted head-banging algorithm to produce maps of general spatial trends free of background noise. Maps of White female lung cancer rates from the new atlas are presented here to illustrate how this innovative combination of maps and graphs permits greater exploration of the underlying mortality data than is possible from previous single-map atlas designs. Published in 1999 by John Wiley & Sons, Ltd. This article is a U.S. Government work and is in the public domain in the United States.

1. INTRODUCTION

Mapping has been an important tool in public health research since John Snow linked the London cholera epidemic to a contaminated water source by mapping the residences of cases.[1] Over the following 100 years, maps of incident cases and death rates have provided aetiologic clues about many diseases, but these maps were limited to either detailed views of a single area or national maps at the state or regional level. It has only been since the 1970s that computer systems have been sufficiently powerful to automate the mapping of small area rates for the entire U.S. Subsequently, mortality atlases have been published for cancer, injuries and other cancer-related causes of death in the U.S.[2–8]

The two primary purposes of a mortality atlas are: (i) to identify specific locations where changes in health policy need to be made, or prevention, screening or treatment programmes started; (ii) to explore more general patterns in the mortality data to generate aetiologic clues for further study. The atlases published to date were successful at identifying disease 'hot spots' for these purposes.[9] For example, follow-up studies to the first cancer atlas uncovered the links

* Correspondence to: Linda W. Pickle, NCI/DCCPS, 6130 Executive Boulevard, MSC 7344, EPN Rm. 313, Bethesda, MD 20892, U.S.A. E-mail: picklel@mail.nih.gov

Published in 1999 by John Wiley & Sons, Ltd.

between shipyard asbestos exposure and lung cancer[10] and snuff dipping and oral cancer.[11] Later, maps of cancer rates over time identified several states where cervical cancer mortality remained high during the 1970s, despite sharp declines elsewhere by that time.[5] Alerted by these maps, these state health departments changed their Medicaid policy to cover pap smears for poor women, and their death rates subsequently declined.

Despite the success of these earlier atlases, their designs did not permit easy exploration of the mortality data beyond identification of the most extreme rates. As part of the project to produce a small-area atlas of mortality for all leading causes of death in the U.S., we examined the ways in which readers extracted information from statistical rate maps.[12] This research led to an innovative combination of maps and graphs that permits greater exploration of the underlying mortality data than was possible from previous single-map atlas designs.

Previous atlas designs comprised a single map for each cause of death[2–4,8] or for each time period.[5–7] Plate 1 is the map of lung cancer death rates among White males for 1950–1969 from the first small-area atlas.[2] The white areas have rates significantly lower than the U.S. rate; dark gold area rates are not significantly different from the U.S. rate; and the gold, orange and red areas are significantly high or in the highest 10 per cent of counties or both. Many important findings came from this map. The National Cancer Institute conducted a number of case-control studies in the highest rate counties (red), and found plausible explanations for the high rates in every area (that is, these areas were not high due to chance). These explanations included unusual smoking habits,[13,14] poor diet,[15,16] exposure to asbestos during World War II shipbuilding,[10,17–19] and arsenic exposure from a smelter plant.[20]

This map design was successful at pointing to specific locations with unusually high death rates, places where epidemiologists found a sufficient number of cases to study, which led to important discoveries about the causes of lung cancer. However, this design does *not* permit easy exploration of the patterns in the data, primarily because most of the counties with rates not significantly different from the U.S. rate are grouped into a single colour category. For example, we cannot tell from this map whether there is a gradient of decreasing rates from the Louisiana Gulf Coast northward because so many places in this region have non-significant rates.

We believed that we could improve upon this design for the new *Atlas of United States Mortality*[21] to permit this kind of data exploration. A literature review revealed conflicting advice from experts as to what design was best for maps of rates.[22–26] Because our Office of Research and Methodology has had an ongoing research programme of cognitive studies in the area of questionnaire design, we decided to draw on this in-house expertise and start a research programme in statistical rate map design. In the following sections, we summarize this research, then discuss how these findings influenced the new *Atlas* design, and finally provide an example to illustrate how this design can be used to explore the data.

2. RESEARCH IN STATISTICAL RATE MAP READING

2.1. Cognitive Research

Early focus groups with epidemiologists identified three types of questions that a typical mortality atlas reader wanted to answer, following Bertin's[27] task classification for graphics:

1. Rate readout task – what is the approximate rate in a specified place?

2. Pattern recognition task – are there regional patterns or clusters of high rates in the data?
3. Pattern comparison task – how do patterns in the maps differ by cause of death, race or gender?

Because the audience for a national mortality atlas is a diverse group of public health administrators and researchers, we wanted to design the map and page layout to permit all of these questions to be answered. In addition to in-house focus groups, we collaborated with a number of university researchers to conduct cognitive experiments in statistical rate map reading. The choice of basic map style was examined, as were the design of the legend, the colour scheme, and the method to indicate unreliable rates.

We considered both classed and unclassed choropleth (area shaded) maps, symbol (dot) maps, and isopleth (contour line) maps as the basic map style for the *Atlas*. Epidemiologists liked the classed choropleth maps and used them most accurately.[28,29] People with some training in cartography preferred the more complex map designs,[30] but they were equally accurate on all types. A separate study by Lewandowsky *et al.*[31] showed that novice map readers identified rate clusters more consistently using monochromatic choropleth maps. Therefore the choropleth map style was used for the *Atlas*.

A study of various legend designs found that the epidemiologists wanted to read the legend quickly and get right on to the more interesting map content.[32] Innovative legend styles were rejected as requiring too much time to understand. A standard fixed box, vertically oriented legend was used for the *Atlas*.

In several studies of map colour schemes, consistent with the recommendations of cartographers, we found that very distinct colours (for example, a rainbow palette) were best for reading a single rate off the map, but that a colour gradient facilitated pattern recognition.[31,33] Based on these results, we decided to use a double-ended colour scheme, a compromise between distinct hues and a colour gradient, for maps where both rate readout and pattern recognition questions would be asked. In the double-ended scheme, a gradient of one hue is used for rates higher than some middle value, and a gradient of another hue is used for lower rates. Carswell has shown the impact on map-reading performance of specific colour choices for double-ended schemes.[34] Particularly important is the need to balance the saturation of high and low rate colours and to use a 'conventional' colour ordering.[34] The final Atlas map colors were chosen so that no one colour dominated a map or a page and so that readers with one of several types of colour blindness could still see distinct shades across the colour categories.[35]

In the earlier cancer atlases, areas that were not particularly high or significantly different from the U.S. rate were grouped into a single colour category, regardless of the actual rate. In hindsight, it seemed that this masked important information about local patterns. For example, no individual rate in an area might be significantly different from the U.S. rate, but a group of adjacent small areas with similar rates might constitute an important spatial cluster or indicate a geographic trend. On the other hand, the reader should be warned when reading a single rate from the map if that rate is based on such small numbers that it should be considered unreliable. Studies by MacEachren *et al.*[36] and Lewandowsky and Behrens[37] showed that hatching over the choropleth colour shading allowed readers to separate the information about the level of the rate and its reliability, that is, they could ignore the hatching and find patches of similar colour for the clustering task, but would be warned about unreliability during a rate readout task.

2.2. Statistical Research

In addition to the cognitive research on map design, we also developed a generalized linear mixed effects model that can help clarify patterns in the underlying data.[21] The logarithms of the age- and place-specific rates were modelled as a spline function of age (either linear or cubic depending on the cause of death). The parameter vector for each HSA was assumed to decompose into fixed regional effects and random small area (HSA) effects within that region.

Results from this model provided regional rate estimates with confidence limits, predicted age-specific rates for each HSA, and improved rate variance estimates. The predicted HSA rates were smoothed using a weighted head-banging algorithm[38,39] to highlight the broad spatial patterns in the data. This two-dimensional median-based smoother was weighted by the inverse of each rate's standard error, so that rates based on few deaths were more likely to be smoothed toward neighbouring area rates than those based on large numbers of deaths.

3. FEATURES OF THE NEW ATLAS

The *Atlas of United States Mortality*[21] is a compendium of maps and graphs of age-adjusted and age-specific rates for 18 leading causes of death, which accounted for 83 per cent of all deaths in the U.S. during 1988–1992. All results are shown by sex and race (White and Black). A new geographic unit, the Health Service Area, is used for mapping. These are groups of counties aggregated on the basis of where county residents obtain routine hospital care.[40] This is the first use of mixed effects modelling and weighted smoothing in a mortality atlas and, as far as we know, this is the first atlas design based on cognitive research.

Plates 2 and 3 show the final two-page *Atlas* layout for lung cancer death rates among White females. Unlike previous atlases, there are multiple maps and graphs for each cause of death, in order to answer different questions about the data. The full page map (Plate 2) displays the age-adjusted death rates, directly adjusted to the 1940 standard population. Its larger size allows the reader to discern the colour of any single HSA so as to read off its observed rate (within a range). The legend for this map is doubly labelled, showing both the actual rate range and the comparative mortality ratio (the HSA rate relative to the U.S. rate). The distribution of the 805 rates is shown at the bottom of the page. The rates are categorized according to percentiles of this distribution (cutpoints at 10, 20, 40, 60, 80, and 90 per cent of the cumulative distribution).

The second page of the layout (Plate 3) contains quarter page size maps and a graph. Because of the smaller size, only five colours are used for each graphic. The upper left map indicates whether each HSA's rate is significantly different from the U.S. rate. The remainder of this page consists of results of the statistical modelling. A row plot[41] of regional rates and 95 per cent confidence limits for two representative ages is shown in the upper right. On the lower half of this page are the smoothed age-specific HSA rates for the same two ages. Because the purpose of these maps is to show broad patterns in the data, a monotonic colour gradient is used; the other maps use a double-ended colour scheme to facilitate both rate readout and pattern recognition tasks.

4. EXAMPLE: WHITE FEMALE LUNG CANCER PATTERNS

To illustrate how this design can be used to explore spatial patterns in the data, consider the lung cancer data for White females shown in Plates 2 and 3. There are no obvious distributional

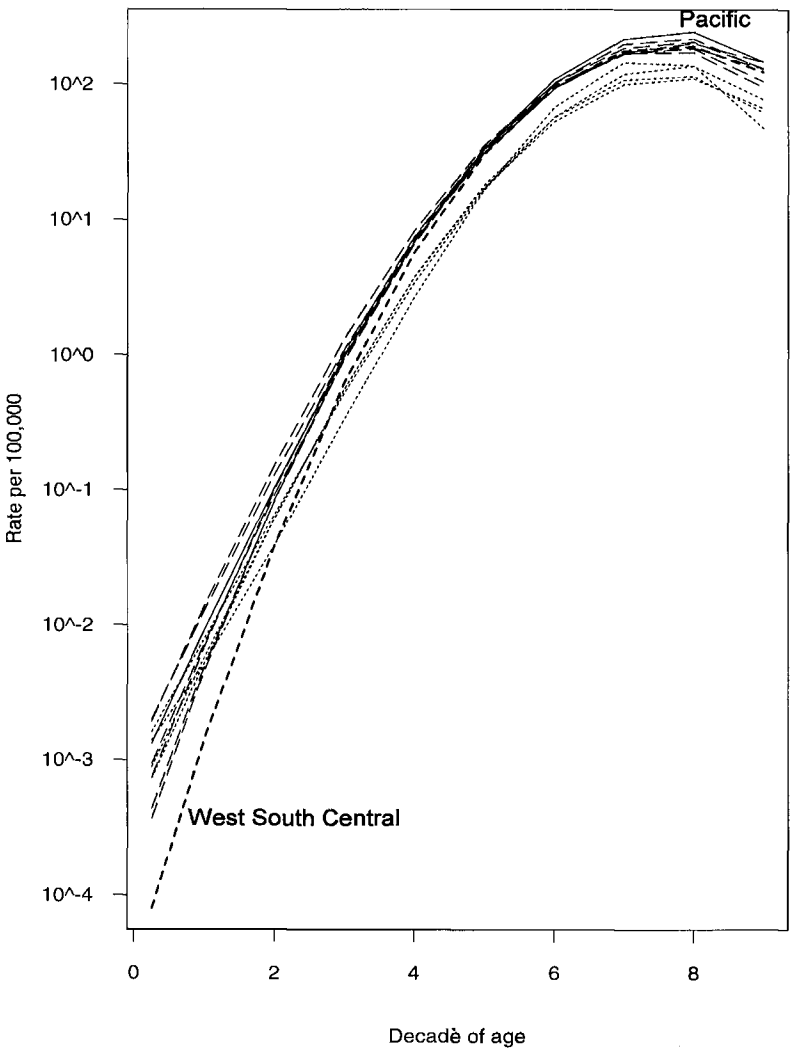
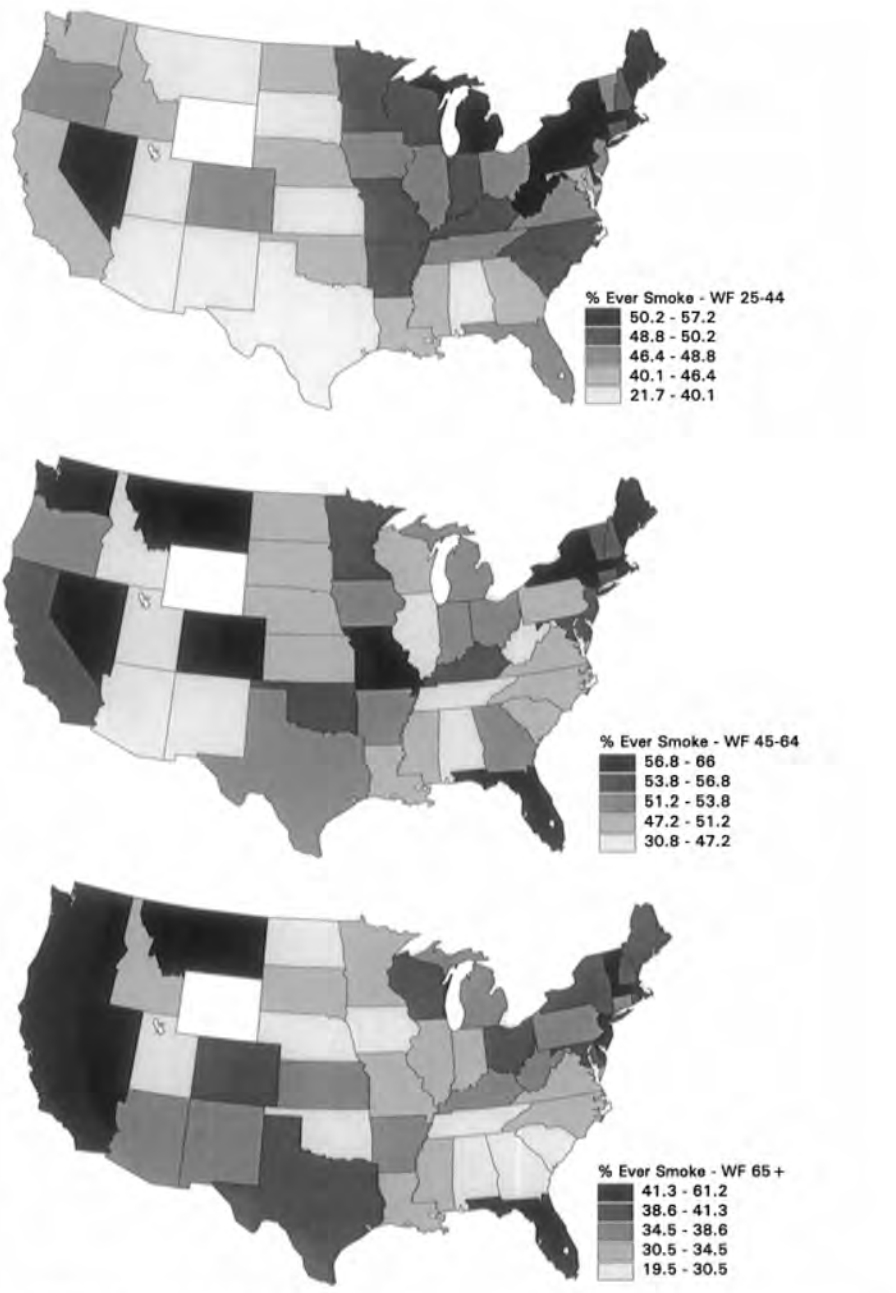


Figure 1. Predicted age-specific rates by region for lung cancer among White females, 1988–1992 (six regions east of Mississippi River shown as dashed lines, Pacific region shown as solid line, other regions shown as dotted lines)

outliers, but there appear to be regional differences in the rates (Plate 2). The age-adjusted map seems to have at least two clusters of high rates – one centred in Kentucky, and one in northern California. Although the California cluster was noted for 1970s data,[5] the apparent cluster for 1988–1992 is larger, now including parts of the Mountain region. The extension of high rates into Nevada, though, is questionable; although technically reliable, the rates are based on relatively few deaths in this sparsely populated region.

Using the results of the mixed effects model, we can examine the regional patterns in more detail. The upper right panel of Plate 3 indicates that both for age 40 and age 70, the 12 regions appear to segregate into two groups. In general, the regions east of the Mississippi River (first 6 of 12 regions) have higher rates, and western regions have lower rates for both ages, with two



Source: CDC Behavioral Risk Factor Surveillance System, 1993.
Note: States categorized into quintiles for each age group. Thus only relative comparisons of patterns can be made across age groups.

Figure 2. Proportion of White women who ever smoked cigarettes by state for ages 25–44, 45–64, 65 and over

Published in 1999 by John Wiley & Sons, Ltd.

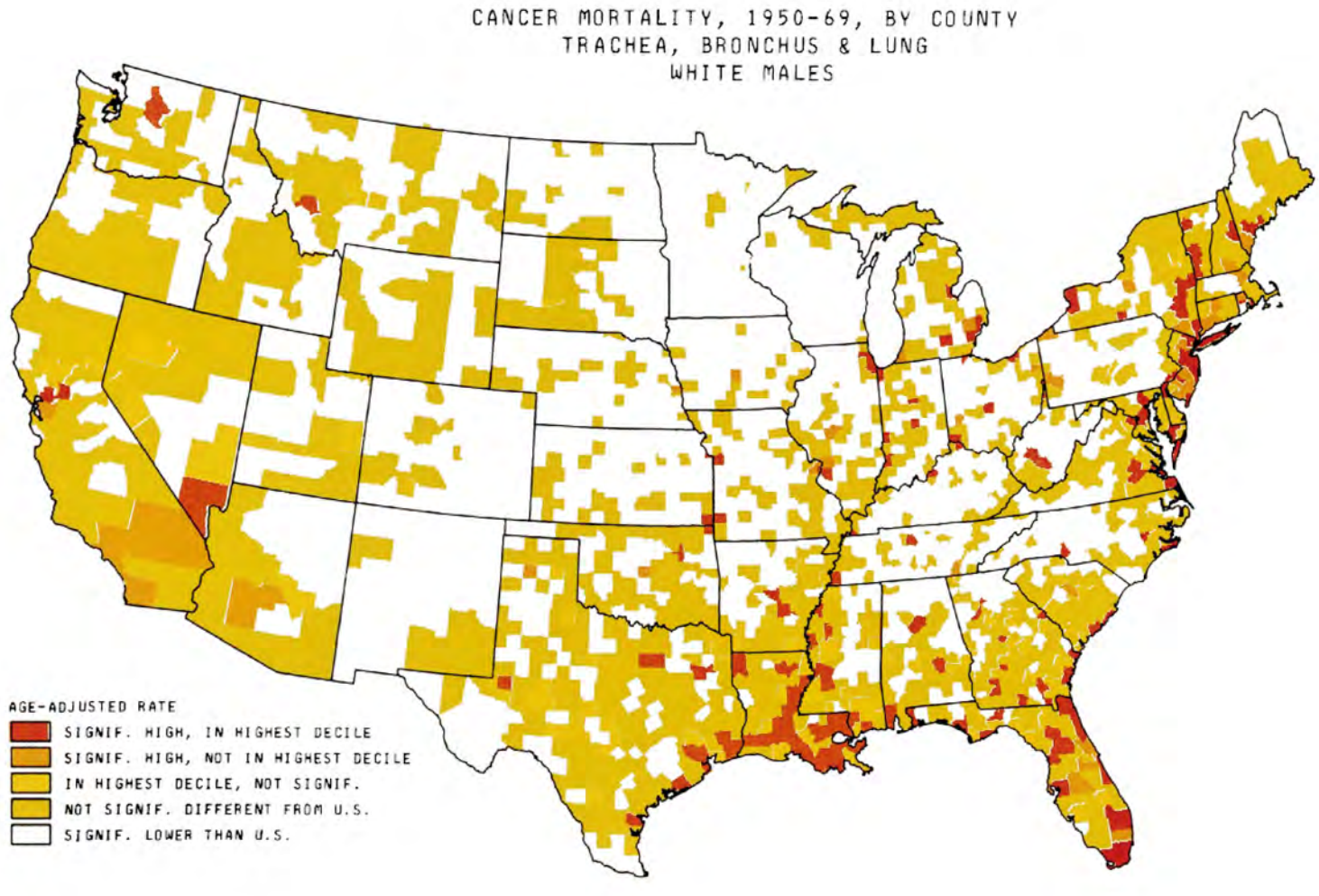


Plate 1. Lung cancer mortality rates for 1950–1969 among White males[2]

Published in 1999 by John Wiley & Sons, Ltd.

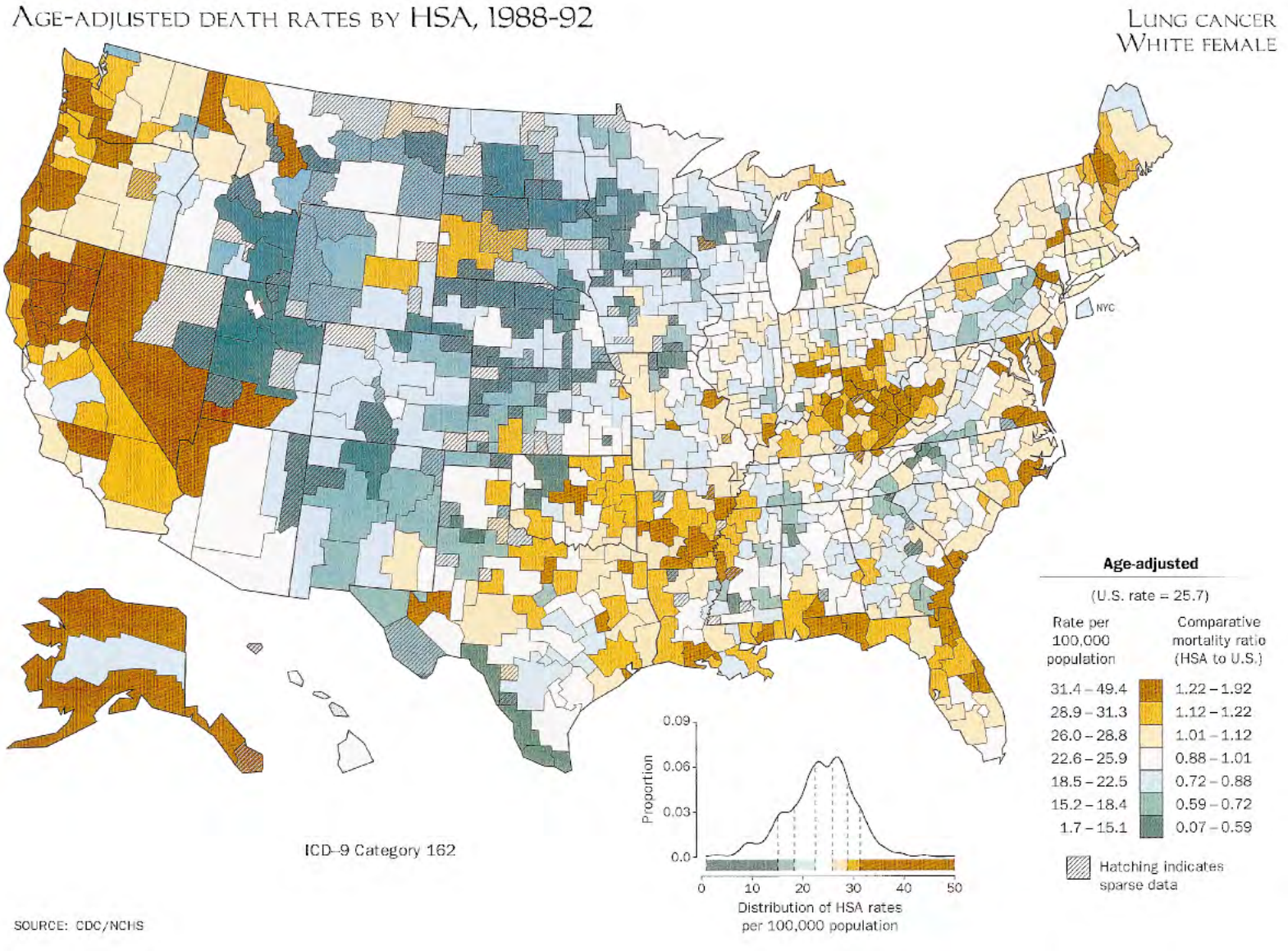


Plate 2. Lung cancer mortality rates 1988–1992 among White females,[21] age-adjusted rates

Published in 1999 by John Wiley & Sons, Ltd.

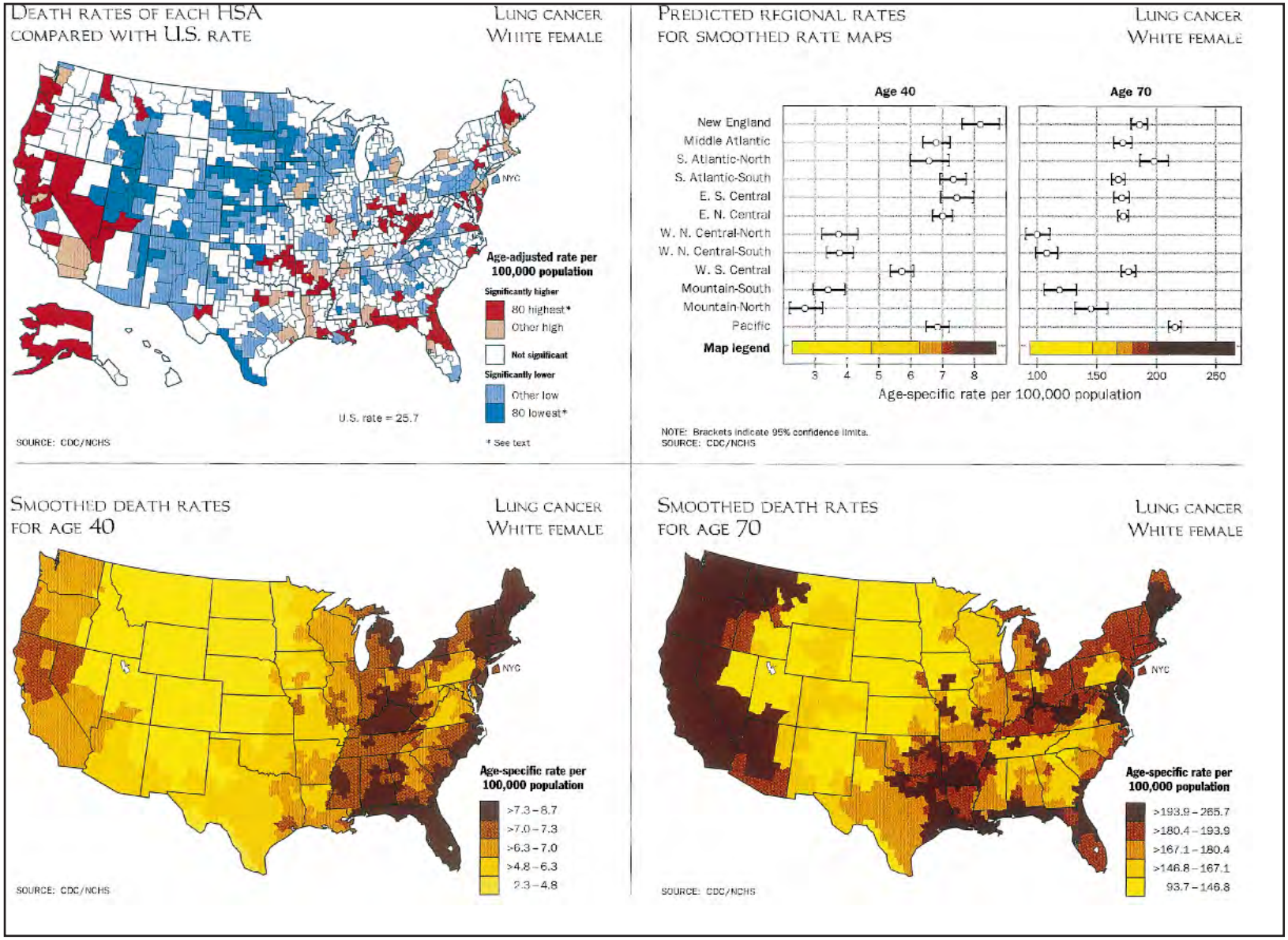


Plate 3. Lung cancer mortality rates for 1988–1992 among White females.[21]
Top left: significance of each HSA rate compared to the U.S. rate.
Top right: predicted regional rates and 95 per cent confidence intervals.
Lower left and right: smoothed predicted HSA rates for age 40 and age 70

exceptions. First, the West South Central region (TX, OK, LA, AR) has a rate intermediate between the eastern and western rates for the younger age, but a higher rate for the older group. Second, the Pacific region has high rates resembling those of the eastern regions, and in fact has the highest regional rate for age 70.

This snapshot of two representative ages can be compared to the entire set of predicted regional age-specific rate curves as shown in Figure 1. Beyond approximately age 25 the regional curves do cluster into high- and low-rate groups. The West South Central region crosses over from the low-rate to the high-rate group between ages 30 and 50. The Pacific region emerges as the highest rate region by age 60. Because of the stability of the regional rates and the regularity of the age-specific rate curve for each cause of death, the row plot for only two ages does give a fair representation of the entire set of curves.

The smoothed age-specific rate maps (bottom, Plate 3) show clear differences in geographic patterns by age. Although rates for the older group are much higher than those for the middle age group, rates east of the Mississippi River are relatively higher than those in the west for age 40 while the reverse is true for age 70. A comparison of these patterns with those in Plate 2 reveals that the apparent clusters of high rates on the age-adjusted map are influenced by different age groups; both age groups contribute to the high age-adjusted rate cluster centred in Kentucky, while the high age-adjusted rates on the West Coast are more influenced by the older group. These differences in mortality patterns by age are consistent with either a cohort effect of decreasing exposure by age to some widespread risk factor for this disease, or the presence of effective prevention programmes targeted at younger residents in a few regions. Maps of the proportion of White women in each state who ever smoked cigarettes (Figure 2) support the former hypothesis; the proportion of smokers is relatively high in the Pacific region only for the oldest group of women. If in fact these maps represent regional differences in the smoking habits of White women by age cohort, we would expect the Pacific lung cancer excess to disappear over the next generation.

5. CONCLUSION

In conclusion, results from the mixed effects model and from the cognitive experiments combine to allow us to explore the geographic patterns of mortality at both larger and smaller scales than was ever possible before – that is, both regional and small area patterns of age-specific and age-adjusted rates. As illustrated by the lung cancer example, differences in patterns across these different scales can provide clues about the reasons for these patterns. In addition, separation of different types of information into several maps allows the reader to inspect the data in a way appropriate for the question being asked. For example, spatial patterns of the rates may be examined regardless of the significance of individual rates, but significance information is available if needed. Supplemental graphics, that is, the density plot and row plot, provide statistical information not available from the maps themselves. Together, these maps and graphs provide more information than ever before available to epidemiologists who are interested in the geographic patterns of disease.

APPENDIX

States included in each region:

1. New England: Maine, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island.
2. Middle Atlantic: New York, Pennsylvania, New Jersey.

3. South Atlantic-North: Delaware, Maryland, Virginia, West Virginia, District of Columbia.
4. South Atlantic-South: North Carolina, South Carolina, Georgia, Florida.
5. East South Central: Kentucky, Tennessee, Mississippi, Alabama.
6. East North Central: Wisconsin, Michigan, Illinois, Indiana, Ohio.
7. West North Central-North: North Dakota, South Dakota, Minnesota.
8. West North Central-South: Nebraska, Iowa, Kansas, Missouri.
9. West South Central: Texas, Oklahoma, Arkansas, Louisiana.
10. Mountain-North: Idaho, Montana, Wyoming.
11. Mountain-South: Nevada, Utah, Colorado, Arizona, New Mexico.
12. Pacific: Washington, Oregon, California.

Because of sparse populations, the West North Central and Mountain regions were not further subdivided for Blacks and Alaska and Hawaii data were not modelled.

REFERENCES

1. Snow, J. *On the Mode of Communication of Cholera*, 2nd edn, The Commonwealth Fund, New York, 1855.
2. Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J. and Fraumeni, J. F. Jr. *Atlas of Cancer Mortality for U.S. Counties*: *1950–1969*, USGPO, DHEW publ. No. (NIH) 75–780, Washington, DC, 1975.
3. Mason T. J., McKay, F. W., Hoover, R., Blot, W. J. and Fraumeni, J. F. Jr. *Atlas of Cancer Mortality among U.S. Nonwhites*: *1950–1969*, USGPO, DHEW publ. No. (NIH) 76–1204, Washington, DC, 1976.
4. Mason T. J., Fraumeni, J. F. Jr., Hoover, R. and Blot, W. J. *An Atlas of Mortality from Selected Diseases*, USGPO, DHHS publ. No. (NIH) 81–2397, Washington, DC, 1981.
5. Pickle, L. W., Mason, T. J., Howard, N., Hoover, R. and Fraumeni, J. F. Jr. *Atlas of U.S. Cancer Mortality among Whites*: *1950–1980*, USGPO, DHHS publ. No. (NIH) 87–2900, Washington, DC, 1987.
6. Pickle, L. W., Mason, T. J., Howard, N., Hoover, R. and Fraumeni, J. F. Jr. *Atlas of U.S. Cancer Mortality among Nonwhites*: *1950–1980*, USGPO, DHHS publ. No. (NIH) 90–1582, Washington, DC, 1990.
7. Riggan W. B., Creason, J. P., Nelson, W. C., Manton, K. G., Woodbury, M. A., Stallard, E., Pellom, A. C. and Beaubier, J., *U.S. Cancer Mortality Rates and Trends*, *1950–1979*, *Vol. IV*: *Maps*, USEPA, Research Triangle Park, NC, 1987.
8. Devine O. J., Annest, J. L., Kirk, M. L., Holmgreen, P., Emrich, S. S., Rosenberg, M. L., Houk, V. N. and Roper, W. L. *Injury Mortality Atlas of the United States*, *1979–1987*, DHHS, PHS, Atlanta, GA, 1991.
9. Hoover, R., Mason, T. J., McKay, F.W. and Fraumeni, J. F. Jr. 'Cancer by county: New resource for etiologic clues', *Science*, **189**, 1005–1007 (1975).
10. Blot, W. J., Harrington, M., Toledo, A., Hoover, R., Heath, C. W. Jr. and Fraumeni, J. F. Jr. 'Lung cancer after employment in shipyards during World War II', *New England Journal of Medicine*, **299**, 620–624, (1978).
11. Winn, D. M., Blot, W. J., Shy, C. M., Pickle, L. W., Toledo, A. and Fraumeni, J. F. Jr. 'Snuff dipping and oral cancer among women in the southern United States', *New England Journal of Medicine*, **304**, 745–749 (1981).
12. Pickle, L. W. and Herrmann, D. J. (eds). *Cognitive Aspects of Statistical Mapping*, NCHS Working Paper Series, No. 18., National Center for Health Statistics, Hyattsville, MD, 1995.
13. Correa, P., Pickle, L. W., Fontham, E., Dalager, N., Lin, Y., Haenszel, W. and Johnson, W. D. 'The causes of lung cancer in Louisiana', *in* Mizell, M. and Correa, P. (eds). *Lung Cancer*: *Causes and Prevention*, Verlag-Chemie International, Deerfield Beach, FL, 1984.
14. Wilcox, H. B., Schoenberg, J. B., Mason, T. J., Bill, J. S. and Stemhagen, A. 'Smoking and lung cancer: risk as a function of cigarette tar content', *Preventive Medicine*, **17**, 263–272 (1988).
15. Fontham, E. T. H., Pickle, L. W., Haenszel, W., Correa, P., Lin, Y. and Falk, R. T. 'Dietary vitamins A and C and lung cancer risk in Louisiana', *Cancer*, **62**, 2267–2273 (1988).

16. Ziegler, R. G., Mason, T. J., Stemhagen, A., Hoover, R., Schoenberg, J. B., Gridley, G., Virgo, P. W. and Fraumeni, J. F. Jr. 'Carotenoid intake, vegetables, and the risk of lung cancer among white men in New Jersey', *American Journal of Epidemiology*, **123**, 1080–1093 (1986).
17. Schoenberg, J. B., Stemhagen, A., Mason, T. J., Patterson, J., Bill, J. and Altman, R. 'Occupation and lung cancer risk among New Jersey white males', *Journal of the National Cancer Institute*, **79**, 13–21 (1987).
18. Blot, W. J., Morris, L. E., Stroube, R., Tagnon, I. and Fraumeni, J. F. Jr. 'Lung and laryngeal cancers in relation to shipyard employment in coastal Virginia', *Journal of the National Cancer Institute*, **65**, 571–575 (1980).
19. Blot, W. J., Davies, J. E., Brown, L. M., Nordwall, C. W., Buiatti, E., Ng, A. and Fraumeni, J. F. Jr. 'Occupation and the high risk of lung cancer in northeast Florida', *Cancer*, **50**, 364–371 (1982).
20. Lee, A. M. and Fraumeni, J. F. Jr. 'Arsenic and respiratory cancer in man: an occupational study', *Journal of the National Cancer Institute*, **42**, 1045–1052 (1969).
21. Pickle, L. W., Mungiole, M., Jones, G. K. and White, A. A. *Atlas of United States Mortality*, National Center for Health Statistics, Hyattsville, MD, 1996.
22. Dent, B. D. *Cartography*: *Thematic Map Design*, Wm. C. Brown, Dubuque, Iowa, 1993.
23. Monmonier, M. *Mapping It Out*, University of Chicago Press, Chicago, 1993.
24. Tukey, J. W. 'Statistical mapping: what should not be plotted', in *Proceedings of the 1976 Workshop on Automated Cartography and Epidemiology*, DHEW Publ. No. (PHS) 79–1254, Hyattsville, MD, 1979.
25. Cleveland, W. S. and McGill, R. 'Graphical perception: theory, experimentation, and application to the development of graphical methods', *Journal of the American Statistical Association*, **79**, 531–534 (1984).
26. Tufte, E. R. 'Design of a cancer atlas', National Center for Health Statistics contract report, 1993.
27. Bertin J. *Semiology of Graphics*: *Diagrams*, *Networks*, *Maps*, University of Wisconsin, Madison, WI, 1983.
28. Maher, R. J. 'The interpretation of statistical maps as a function of the map reader's profession', *in* Pickle, L. W. and Herrmann, D. J. (eds.), *Cognitive Aspects of Statistical Mapping*, NCHS Working Paper Series, No. 18, National Center for Health Statistics, Hyattsville, MD, 1995, pp. 249–274.
29. Pickle, L. W., Herrmann, D., Kerwin, J., Croner, C. M. and White, A. A. 'The impact of statistical graphic design on interpretation of disease rate maps', Proceedings of the Statistical Graphics Section of the 1993 Annual Meeting of the American Statistical Association, 1994, pp. 111–116.
30. White, A. A., Pickle, L. W., Herrmann, D. J., Croner, C. M. and Wilson, B. F. 'Map design preferences associated with professional discipline', Proceedings of the Statistical Graphics Section of the 1994 Annual Meeting of the American Statistical Association, 1995, pp. 54–59.
31. Lewandowsky, S., Herrmann, D. J., Behrens, J. T., Li, S.-C., Pickle, L. and Jobe, J. B. 'Perception of clusters in statistical maps', *Applied Cognitive Psychology*, **7**, 533–551 (1993).
32. Pickle, L. W., Hermann, D. and Wilson, B. 'A legendary study of statistical map reading: the cognitive effectiveness of statistical map legends', *in* Pickle, L. W. and Herrmann, D. J. (eds.), *Cognitive Aspects of Statistical Mapping*, NCHS Working Paper Series, No. 18, National Center for Health Statistics, Hyattsville, MD, 1995, pp. 233–248.
33. Hastie, R., Hammerle, O., Kerwin, J., Croner, C. M. and Hermann, D. J. 'Human performance reading statistical maps', *Journal of Experimental Psychology*: *Applied*, **2**, 3–16 (1996).
34. Carswell, C. M., Kinslow, H. S., Pickle, L. W. and Herrmann, D. 'Using colour to represent magnitude in statistical maps: the case for double-ended scales', *in* Pickle, L. W. and Herrmann, D. J. (eds.), *Cognitive Aspects of Statistical Mapping*, NCHS Working Paper Series, No. 18, National Center for Health Statistics, Hyattsville, MD, 1995, pp. 201–228.
35. Brewer, C. A., MacEachren, A. M., Pickle, L. W. and Herrmann, D. 'Mapping mortality: evaluating colour schemes for choropleth maps', *Annals of the Association of American Geographers*, **87**(3), 411–438 (1997).
36. MacEachren, A. M., Brewer, C. A. and Pickle, L. W. 'Visualizing georeferenced data: representing reliability of health statistics', *Environment and Planning A*, **30**, 1547–1561 (1998).
37. Lewandowsky, S. and Behrens, J. T. 'Accuracy of cluster detection in mortality maps', *in* Pickle, L. W. and Herrmann, D. J. (eds.), *Cognitive Aspects of Statistical Mapping*, NCHS Working Paper Series, No. 18, National Center for Health Statistics, Hyattsville, MD, 1995, pp. 133–148.

38. Mungiole, M., Pickle, L. W. and Simonson, K. H. 'Application of a weighted head-banging algorithm to mortality data maps', *Statistics in Medicine*, **18**, 3201–3209 (1999).
39. Hansen, K. M. 'Head-banging: robust smoothing in the plane', *IEEE Transactions on Geoscience and Remote Sensing*, **29**, 369–378 (1991).
40. Makuc, D. M., Haglund, B. and Ingram, D. D., Kleinman, J. C. and Feldman, J. J. 'Health service areas for the United States', Vital Health Statistics 2(112), National Center for Health Statistics, Hyattsville, MD, 1991.
41. Carr, D. 'Converting plots to tables', Technical report no. 101, Center for Computational Statistics, George Mason University, Fairfax, VA, 1994.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3221–3234 (1999)

ALL MAPS OF PARAMETER ESTIMATES ARE MISLEADING

ANDREW GELMAN[1] AND PHILLIP N. PRICE[2] *

[1] *Department of Statistics, Columbia University, 618 Mathematics Building, New York, New York 10027, U.S.A.*
[2] *Lawrence Berkeley National Laboratory, LBNL 90-3058, Berkeley, California 94720, U.S.A.*

SUMMARY

Maps are frequently used to display spatial distributions of parameters of interest, such as cancer rates or average pollutant concentrations by county. It is well known that plotting observed rates can have serious drawbacks when sample sizes vary by area, since very high (and low) observed rates are found disproportionately in poorly-sampled areas. Unfortunately, adjusting the observed rates to account for the effects of small-sample noise can introduce an opposite effect, in which the highest adjusted rates tend to be found disproportionately in well-sampled areas. In either case, the maps can be difficult to interpret because the display of spatial variation in the underlying parameters of interest is confounded with spatial variation in sample sizes. As a result, spatial patterns occur in adjusted rates even if there is no spatial structure in the underlying parameters of interest, and adjusted rates tend to look too uniform in areas with little data. We introduce two models (normal and Poisson) in which parameters of interest have *no* spatial patterns, and demonstrate the existence of spatial artefacts in inference from these models. We also discuss spatial models and the extent to which they are subject to the same artefacts. We present examples from Bayesian modelling, but, as we explain, the artefacts occur generally. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

1.1. Background

When a spatially-varying parameter of interest is subject to substantial uncertainty, maps of predicted values can differ in important and systematic ways from the spatial distribution of true values. A standard method for correcting for these artefacts – Bayes shrinkage estimation – introduces new and opposite artefacts of its own.

We will illustrate this point with generic statistical models, but it is helpful to keep a specific example in mind. Consider the mapping of cancer mortality rates by county in the United States. Much of the variation in observed cancer death rates by county is attributable to statistical noise due to the small number of (observed and expected) cancer deaths in low-population counties. Because of this stochastic noise, a disproportionate fraction of low-population counties are observed to have extremely high (or low) cancer rates when compared to typical counties in the United States. Thus

* Correspondence to: Phillip N. Price, Lawrence Berkeley National Laboratory, LBNL 90-3058, Berkeley, California 94720, U.S.A. E-mail: pnprice@lbl.gov

Contract/grant sponsor: U.S. National Science Foundation
Contract/grant numbers: DMS-9404305, SBR-9708424, DMS-9457824
Contract/grant sponsor: U.S. Department of Energy
Contract/grant numbers: DE-AC03-76SF00098

CCC 0277–6715/99/233221–14$17.50

when counties with very high observed rates are highlighted on a map of the U.S., almost all of the highlighted counties are low-population counties.[1, 2] Since the Central and Western U.S. contain a great many such counties, a much higher fraction of counties in the Central and Western U.S. is highlighted than in the rest of the country.

Manton *et al.*[1] and Riggan *et al.*[3] use a Bayesian procedure to estimate underlying cancer rates by county. This procedure is now common, with minor variations, for U.S. cancer maps.[4, 5] The posterior mean estimate for each county is a compromise between the observed county cancer mortality rate and the mean cancer mortality rate for the entire U.S. (or for a region of the U.S.[5]), with the relative weighting of these rates being dependent on the county population. The estimated underlying cancer death rate for a high-population county with a given observed rate is close to the observed value, and this estimate has a small standard error. A low-population county with the same observed rate has a posterior mean somewhere between the observed rate and the U.S. mean rate, with a larger standard error.

Manton *et al.*[1] quite reasonably suggest that the posterior mean estimates are more appropriate for mapping than are the observed death rates, since the observed rates are subject to systematic effects related to county population, and since Bayes and empirical Bayes methods tend to yield more accurate predictions than do raw rates.[6–8] Unfortunately, the posterior means are subject to a similar type of systematic artefact related to county population, but in the opposite direction, as we will show. (Similar problems with the ensemble of posterior mean estimates are noted by Louis.[9]) We also show that most other mapping methods have artefacts associated with populations or sample sizes.

We quantify these artefacts in this paper, using the examples of standard models for continuous and discrete data to demonstrate that maps of point estimates can introduce spurious spatial patterns. This occurs even when the model being fit is appropriate, and even when there is *no* underlying spatial structure in the parameter of interest. In Section 2 we consider an example with normally-distributed parameters and measurements. In Section 3, we examine a Poisson/gamma model with parameters taken from cancer data. In Section 4, we discuss the occurrence of artefacts in fitting spatial models to data that *do* have underlying spatial structure.

1.2. Theoretical approach to examining statistical artefacts

If one fits a statistical model that is inappropriate to the data being analysed, then inferences will be incorrect and maps of predictions might well show spurious spatial patterns. This is *not* what we mean by 'spatial artefacts' in this paper. Instead, we consider a spurious spatial pattern to be an 'artefact' if it occurs even when inferences are based on the *correct* statistical model.

We analyse mapping artefacts in the context of a theoretical model with *no* spatial effects. This approach allows us to illustrate our points with simple and easily interpreted statistical models, and makes it easy to see the effects of the artefacts in our sample maps since any apparent spatial pattern is an artefact. As we discuss in Section 4, the same sorts of artefacts occur when the parameter of interest varies spatially, even if the correct spatial model is fit.

Under our model, each of J counties, $j = 1, \ldots, J$, has an unknown parameter θ_j. The ensemble of parameters, $\{\theta_1, \ldots, \theta_J\}$, follows some distribution, $p(\theta_j)$, assumed known. In each county j, we have n_j independent measurements y_{ij}, $i = 1, \ldots, n_j$, with a known sampling distribution: $y_{ij}|\theta_j \sim p(y_{ij}|\theta_j)$. We further assume, in this theoretical model, that the sample sizes n_j are statistically independent of the true parameter values θ_j (so that the values of n_j do not convey information about the θ_j's), and that the parameters θ_j are spatially uncorrelated.

We now suppose that a statistical analysis is performed and then a map is drawn to indicate the estimated value of θ_j in each county. Although this paper applies to parameter mapping in general, we will focus on maps that highlight only the counties with the highest point estimates, so that we can use black and white maps as illustrations. Colour or greyscale maps would manifest the same artefacts – for example, using a colour map to search for 'hot spots' of a parameter would be equivalent to looking at the highlighted counties in our black and white maps.

In our analysis, we ignore the difference between (a) highlighting the top x per cent of counties and (b) highlighting the counties that exceed a threshold that, in expectation, exceeds all but x per cent of the counties. In practice, both procedures are used.[1, 10] The two procedures give essentially the same result. For example, there is little difference between highlighting 27 out of 274 counties or using a fixed threshold so that the expected number of counties highlighted is 27·4. For mathematical simplicity, we consider procedure (b) in this paper.

If the map of extreme values is simply based on point estimates $\hat{\theta}_j$, this means that some threshold c is set so that the counties j for which $\hat{\theta}_j > c$ are highlighted. More generally, if one attempts to adjust for sample size then a function $h(\cdot,\cdot)$ is chosen and a threshold c is set so that the counties for which $h(y_j, n_j) > c$ are highlighted. The main point of this paper is that, for most mapping methods – that is, for most choices of $h(\cdot,\cdot)$ – the probability that a county is highlighted, $\Pr(h(y_j, n_j) > c | n_j)$, depends on the sample size n_j, so that the map of highlighted counties will display patterns based on the sample sizes.

2. CONTINUOUS MEASUREMENTS

We first work out the basic results for the relatively simple problem of continuous measurements with normally-distributed errors. For counties $j = 1, \ldots, J$, let θ_j be the true value of a parameter in county j. We assume that the true values of the county parameters, θ_j, follow a normal distribution:

$$\theta_j \sim \mathrm{N}(\mu, \tau^2). \tag{1}$$

By assuming the county parameters follow a common distribution, we are *not* assuming that the counties are identical – that would correspond to $\tau = 0$. The data from each county constitute n_j independent, identically distributed measurements

$$y_{ij} | \theta_j \sim \mathrm{N}(\theta_j, \sigma^2). \tag{2}$$

2.1. Problems with mapping the sample means

The direct estimate of each θ_j is the observed county mean, which we label y_j. It is well known that, if the n_j's vary, the procedure of selecting the counties with the highest observed means tends to yield counties with few observations; we will quantify this artefact. An observed county mean y_j based on n_j observations is distributed as

$$y_j \sim \mathrm{N}(\mu, \tau^2 + \sigma^2/n_j). \tag{3}$$

Under our model, the probability that the observed mean y_j exceeds a threshold c, for a county with sample size n_j, is

$$\Pr(y_j > c | n_j) = \Phi\left[\frac{\mu - c}{\tau}\left(1 + \frac{\sigma^2}{n_j \tau^2}\right)^{-1/2}\right]. \tag{4}$$

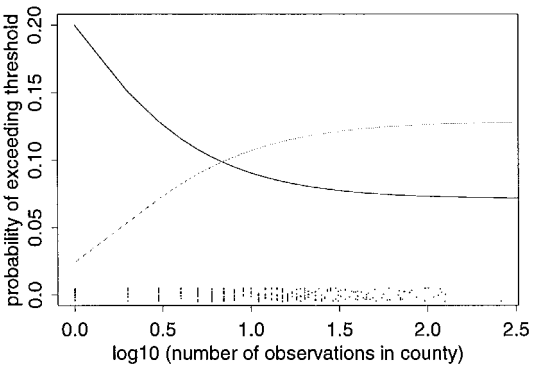


Fig. 1. Solid line: probability that an observed county mean, y_j, will exceed a specified cut-off point, c_1. Dotted line: probability that the posterior mean estimate for a county, $E(\theta_j|y_j)$, will exceed a specified cut-off point, c_2. Both lines are plotted as a function of the log (base 10) of n_j, the number of observations in the county. Each cut-off point is set to catch an average of 10 per cent of the counties. Curves are derived from the values of n_j in the radon data structure and from the variance ratio $\tau^2/\sigma^2 = 0{\cdot}49$ estimated from the radon data. The points at the bottom of the figure show the 274 values of $\log_{10} n_j$; they are jittered (see Chambers *et al.*[12]) so that duplicate values are visible

For any given threshold c, one can compute the expected number of counties that will be shaded under the model by summing the probabilities (4), for a given set of n_j's.

If the threshold c is to be set so that some small fraction of counties (for example, 5, 10, or 20 per cent) is expected to exceed it, then c will almost certainly be larger than the grand mean, μ, and the probability of exceeding it is a decreasing function of n_j. The variation of this probability with n_j depends on both the variance ratio σ^2/τ^2 and the value of c, which itself depends on the distribution of the J values of n_j.

To illustrate, we use the example of home radon levels in the mid-Atlantic region of the U.S., which comprises 277 counties, including some independent cities in Virginia. In this region, the Environmental Protection Agency and the state health departments randomly sampled 5677 homes;[11] three of the counties had no homes surveyed, and of the remaining counties, the number n_j of homes surveyed ranged from 1 to 261. The measurements y_{ij} are the natural logarithms of the measured radon levels, and the parameter θ_j is the average log radon level in county j (that is, the log geometric mean radon measurement that would be obtained if every home in the county were to be measured). We fit a hierarchical normal model[13, 14] to these data and obtained estimates of 1·0 and 0·7 for the within- and between-county standard deviations, σ and τ, respectively.

We study the artefacts created by the mapping procedure for the radon example by working out what would happen if the hierarchical normal-normal model were true, with hyperparameter values $\sigma = 1{\cdot}0$ and $\tau = 0{\cdot}7$. That is, we construct a model in which the statistical distribution of county radon levels is similar to that from the actual data, but in which (unlike the actual radon data) the county parameters are distributed randomly, with *no* spatial correlation. Under this model and the given set of 274 values of n_j, the cut-off value to highlight the top 10 per cent of counties is $c_1 = \mu + 1{\cdot}468\tau$.

The solid line on Figure 1 shows the probability that any given county mean y_j will exceed c_1, as a function of $\log_{10} n_j$. The points at the bottom of the figure show the values of $\log_{10} n_j$ in the data set. (Ignore the dotted line on the figure for now.) Counties with fewer than about six measurements are much more likely to exceed the threshold than are more heavily sampled

counties. This statistical artefact manifests itself as a spatial artefact in a map of county means, because the sample sizes themselves vary spatially.

2.2. Problems with mapping the posterior point estimates

It has been suggested[1] that one should map the county posterior mean estimates, $E(\theta_j|y_j,n_j)$, to avoid the artefact discussed above. Unfortunately, mapping county posterior means or highlighting the counties with highest posterior means leads to new problems. Under the normal model above, the posterior mean (and mode) estimate for a county is

$$E(\theta_j|y_j,n_j)=\frac{\dfrac{1}{\tau^2}\mu+\dfrac{n_j}{\sigma^2}y_j}{\dfrac{1}{\tau^2}+\dfrac{n_j}{\sigma^2}}. \tag{5}$$

Averaging over the marginal distribution of y_j, we find that, for a county with sample size n_j, the probability that $E(\theta|y_j,n_j)$ exceeds a fixed value c is

$$\begin{aligned}\Pr(E(\theta_j|y_j,n_j)>c|n_j)&=\Pr\left(y_j>\frac{\sigma^2}{n_j}\left[\left(\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}\right)c-\frac{1}{\tau^2}\mu\right]\middle|\,n_j\right)\\&=\Phi\left[\frac{\mu-c}{\tau}\left(1+\frac{\sigma^2}{n_j\tau^2}\right)^{1/2}\right]\end{aligned} \tag{6}$$

an expression which is similar to (4) but is now an increasing, rather than decreasing, function of n_j (assuming $c>\mu$, which will be the case if the cut-off is set so that a small fraction of counties will be highlighted).

Mapping county posterior mean estimates (5) still leads to artefacts related to sample sizes, since $E(\theta_j|y_j,n_j)$ depends on n_j. For example, in the radon data much of West Virginia was sparsely sampled (values of n_j were low), so that a map of the posterior estimates of county means in West Virginia will appear quite uniform even if the true county levels θ_j are highly variable.

Under the assumed model, the threshold c that leads to an expected 10 per cent of the counties being highlighted is $c_2=\mu+1{\cdot}132\tau$, a lower value than the cut-off c_1 for the raw county means, which makes sense since the posterior mean estimates are shrunken towards the grand mean. (We computed c_2 by iteratively trying different values of c until the average value of (6), averaging over the counties j, was 10 per cent.) The dotted line in Figure 1 displays the probability of a county's posterior mean estimate exceeding c_2, as a function of $\log_{10}n_j$. Clearly, a map highlighting the posterior means has a strong artefact in the opposite direction to the map of the observed means; the counties with fewer than six observations are disproportionately *unlikely* to have notably high posterior means.

2.3. Problems with maps based on statistical significance

Other natural methods of mapping extreme counties also suffer from artefacts so that the probability of a county being highlighted depends on the number of observations in the county. For example, one could highlight the counties with the highest posterior probability of exceeding some specified level, $\mu+x\tau$. Under the normal model, this is equivalent to choosing the counties with the highest 'posterior z-scores', $z_j=(E(\theta_j|y_j,n_j)-(\mu+x\tau))/\mathrm{sd}(\theta_j|y_j,n_j)$. The resulting probability that a

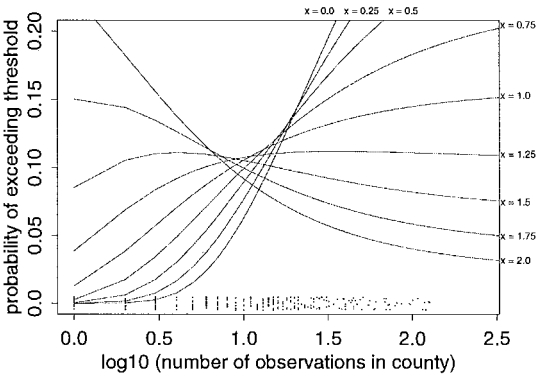


Fig. 2. Probability that a county will be in the top 10 per cent of counties as ranked by $\Pr(\theta_j > \mu + x\tau | y_j)$. Curves shown for $x = 0, 0{\cdot}25, 0{\cdot}5, \ldots, 2{\cdot}0$; each curve is plotted as a function of the log (base 10) of n_j, the number of observations in the county. Curves are derived from the values of n_j in the radon data structure and from the variance ratio $\tau^2/\sigma^2 = 0{\cdot}49$ estimated from the radon data. The points at the bottom of the figure show the 274 values of $\log_{10} n_j$

county with sample size n_j is highlighted is

$$\Pr(z_j > z_c | n_j) = \Phi\left[-\frac{z_c\sigma}{n_j^{1/2}\tau} - x\left(1 + \frac{\sigma^2}{n_j\tau^2}\right)^{1/2}\right] \tag{7}$$

where z_c is the cut-off z-score level set so that 10 per cent (say) of counties are highlighted. Expression (7) is dependent on n_j in a relatively complicated manner; note that z_c can be either positive or negative, depending on x and the data structure.

To illustrate, Figure 2 displays the probability that a county is highlighted, as a function of $\log_{10} n_j$, for each of several values of x, from $x = 0$ (corresponding to selecting the 10 per cent of counties with the highest posterior probability of $\theta_j > \mu$) to $x = 2$ (corresponding to selecting the 10 per cent of counties with the highest posterior probability of $\theta_j > \mu + 2\tau$), for the radon data structure. None of these curves is a constant function of n_j, but for $x = 1{\cdot}5$ the curve is close to flat, corresponding to a mapping procedure that is relatively free of artefacts due to sample sizes.

Should we, then, construct a map based on the rankings of the counties in terms of $\Pr(\theta_j > \mu + 1{\cdot}5\tau | y_j, n_j)$? We think not, because this is not a natural measure or ranking. In fact, the '1·5' depends on the structure of the data and would change if σ/τ or the set of n_j's were changed, so maps of different data (death rates from different cancer types, for instance) would require disparate ranking methods to avoid spatial artefacts. In addition, it is not clear what relevance such a measure as $\Pr(\theta_j > \mu + 1{\cdot}5\tau | y_j, n_j)$ would have to any questions of inherent scientific interest. Using such a measure would reduce artefacts due to sample sizes, but only at the expense of the ease of interpretability that is one of the reasons for producing maps in the first place.

A related approach to weeding out the highly variable small counties is to highlight the counties that are statistically significantly greater than the overall mean – in the normally-distributed case, this would mean $y_j > \mu + 2\sigma/n_j^{1/2}$. This method can be an improvement on merely mapping y_j (see, for example, Tufte,[15] pp. 16–19, who displays maps from Mason *et al.*[16] indicating both extreme values of y_j and statistical significance, and Schlattmann *et al.*[17], who map Bayes estimates indexed by statistical significance). However, as with all the other methods we have considered

so far, maps highlighting statistical significance do not eliminate artefacts based on sample size; if the sample size in a county is extremely large, even a small difference between the county's observed rate y_j and the mean rate μ will be statistically significant, so again this method is more likely to include a high population county than a low-population one with the same true parameter value.[18]

In the normal model, artefacts based on sample size can be eliminated by highlighting the counties for which the quantity $z_{\text{marg}} = (y_j - \mu)/(\tau^2 + \sigma^2/n_j)^{1/2}$, is highest. We label this the *marginal z-score*, because it measures the discrepancy of the county mean y_j with respect to its marginal distribution, averaging over the unknown county parameter θ_j. Under the assumed model, $\Pr(z_{\text{marg}} > c|n_j)$ is just the cumulative standard normal distribution evaluated at c and does not depend on n_j – thus, no artefacts due to sample size. (Incidentally, this works only for continuous data; any discreteness in the distribution of y_j causes the probabilities to vary with n_j.) A map of the extreme values of z_{marg} could be a useful kind of 'standardized residuals' plot. However, such a map still has the same problem as the other proposals mentioned in this section; the mapped values have no direct interpretation as estimates of θ_j. For example, the low-sample-size counties highlighted on such a map will have lower values of θ_j, on average, than the highlighted counties with high sample size.

2.4. Multiple imputation of posterior parameters

An alternative method of producing maps is to multiply impute the vector of posterior parameters. Multiple imputation[19, 20] is a method of accounting for the posterior uncertainty in a vector, $\theta = (\theta_1, \ldots, \theta_J)$ by drawing L simulations of the vector, θ^l, $l = 1, 2, \ldots, L$. This means drawing each vector θ from the posterior distribution $p(\theta|y)$. A map based on one simulated vector of county parameters $\theta^l = (\theta_1^l, \ldots, \theta_J^l)$ represents just one 'possible' reality. A multiple imputation yields several such maps, each based on a different draw of the vector of county parameters.

For example, if the highest counties were of interest, one could highlight on each map the 10 per cent of counties with highest values of θ_j in that simulation draw. Variation from map to map would show posterior uncertainty. Thus, a county for which no information is available would be highlighted on 1/10 of the maps (after all, it *could* be in the top 10 per cent of true county means); a county with many observations and a very high observed value would be highlighted in nearly all the maps; a county with few observations and a very high observed value would be highlighted on more than 1/10, but perhaps not most, of the maps; and so forth.

A multiply-imputed map does not suffer from the artefacts described in the previous sections. More precisely, *if* the model being applied is correct, and a map is made highlighting all counties with imputed θ_j values higher than some cut-off c, *then* the probability that a county is highlighted in any given imputation does not vary with the sample size, n_j. To see why this is so, notice that the probability that county j is highlighted in a single randomly-produced map, given the data y_j from that county, is just the posterior probability $\Pr(\theta_j > c|y_j, n_j)$. The probability that a particular county with sample size n_j is highlighted in a map, obtained by averaging over the marginal distribution of y_j, is

$$\int \Pr(\theta_j > c|y_j, n_j)\, p(y_j|n_j)\, \mathrm{d}y_j = \Pr(\theta_j > c|n_j) \tag{8}$$

which depends only on the distribution of true county parameters, $p(\theta_j)$, and not on the number of observations n_j. (Recall that we have assumed that θ_j and n_j are statistically independent.)

Of course, use of multiple imputation requires the production of multiple maps if one wishes to examine the spatial distribution and uncertainties of quantities of interest; any single map based on multiple imputation gives no indication of which spatial features are due to chance and which are strongly supported by the data, as we will discuss below in the context of multiple imputation of cancer maps.

3. COUNTED DATA

The occurrence of artefacts related to the amount of information in each map unit is a general result, but the details vary with the model and data structure. We illustrate the case of counted data with the Poisson/gamma model, which is commonly used in small area estimation with data such as cancer incidences; similar results would be obtained, with somewhat more computational effort, under the other standard family of models,[8] the Poisson/log-normal. For counties $j = 1, \ldots, J$, let θ_j be the underlying rate parameter, n_j be the population in county j, and

$$y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ is affected} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we label the observed number of incidences in county j as $y_j = \sum_{i=1}^{n_j} y_{ij}$, so that the observed rate for the county is y_j/n_j. It is then standard to model y_j as a Poisson random variable with parameter $n_j\theta_j$. We further assume that the county parameters θ_j follow a gamma(α, β) distribution.

We illustrate with the data structure of the Manton *et al.*[1] example of ten-year kidney/ureter cancer rates in counties of the United States, with n_j equal to county populations, and with $\alpha = 20$ and $\beta = 20/(4{\cdot}65 \times 10^{-5})$. We chose these parameters so that the mean and variance of the gamma(α, β) distribution would approximately match the mean and variance of the county parameters in the Manton *et al.* paper. From this distribution, we draw a 'true' cancer rate for each county. We also assign an 'observed' rate, drawn from the Poisson $(n_j\theta_j)$ distribution for each county. As before, we do not use the data y_j from Manton *et al.*; rather, we model what would happen if the true values county parameters were drawn from a gamma distribution with the (approximately) correct scale and shape but independently of any spatial or other variables.

We consider the effects of highlighting counties based on the raw means, y_j/n_j, or the posterior means, which are given by

$$E(\theta_j | y_j, n_j) = \frac{\alpha + y_j}{\beta + n_j}. \tag{9}$$

As with the normal model, the mapping artefacts depend on the sample sizes (in this case, the populations), n_j, and the distribution of county rates, θ_j.

Figure 3 is the analogy, under the Poisson-gamma model, to Figure 1. The solid line in Figure 3 shows the probability that any given county mean, y_j/n_j, will exceed c_1, as a function of $\log_{10} n_j$, where $c_1 = 11{\cdot}2 \times 10^{-5}$ is the cut-off set so that one expects 10 per cent of the counties to be highlighted. (For any given threshold c, one can compute the expected number of counties that will be shaded under the model by simulation from the gamma and Poisson distributions. We arrived at the value 11·2 by iteratively altering c until the expected proportion of shaded counties was 10 per cent.) The dotted line shows the probability that any given posterior mean, $(\alpha + y_j)/(\beta + n_j)$, will exceed $c_2 = 5{\cdot}0 \times 10^{-5}$, the cut-off set so that one would expect 10 per cent of the counties

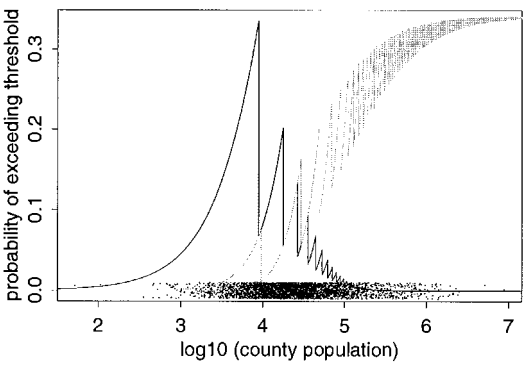


Fig. 3. Solid line: probability that an observed county cancer death rate, y_j, will exceed a specified cut-off point, c_1. Dotted line: probability that the posterior mean estimate for a county, $E(\theta_j|y_j)$, will exceed a specified cut-off point, c_2. Both lines are plotted as a function of the log (base 10) of n_j, the population in the county. Each cut-off point is set to catch an average of 10 per cent of the counties. Curves are derived from the values of n_j in U.S. counties and from the gamma $(20, 4{\cdot}3 \times 10^5)$ distribution fit to the Manton *et al.*[1] data. The points at the bottom of the figure show the 3082 values of $\log_{10} n_j$

to be highlighted under this method. (Recall that the grand mean of the θ_j's is assumed to be $4{\cdot}65 \times 10^{-5}$.) Given the cut-offs c_1 and c_2, we computed probabilities for the solid and dotted lines based on the marginal distribution for y_j, which is negative binomial. The points at the bottom of the figure show the 3082 values of $\log_{10} n_j$ for U.S. counties.

The sawtooth pattern of Figure 3 arises from the discrete nature of the data; for example for a map based on observed rates, a county with n_j in the range $[0, 1/c_1)$ will be highlighted if $y_j \geqslant 1$, whereas if n_j is in the range $[1/c_2, 2/c_1)$, at least two occurrences of cancer are required, and so forth. In addition to the sawtooth pattern, Figures 1 and 3 show different behaviours at the limits of small and large n.

Maps based on observed rates overemphasize the counties with small populations, but maps based on posterior mean have the reverse problem that the more populous counties are more likely to be highlighted. For the model discussed above, the average county population is 80,000, but the expected average population of the highlighted counties is 16,000 if highlighting is based on raw means or 190,000 if highlighting is based on posterior means.

Figure 4 displays the top 10 per cent of counties according to θ_j, for our simulated data; this is equivalent to a random sample of 10 per cent of U.S. counties. Figures 5(a) and (b) display the top 10 per cent of counties according to the observed rates and posterior means, respectively. The patterns – most notably, the presence of many counties from the Mountain and Plains states in the highest 10 per cent based on the observed rates, and the very small fraction of counties in those states in the maps of posterior means – are similar to Figures 1 and 2 of Manton *et al.*,[1] which plot the counties with highest observed rates and posterior means for kidney/ureter cancer death rates. This similarity suggests that many of the spatial patterns in that paper, and in maps of Bayes-smoothed cancer rates in general, are artefactual.

As in the previous example, we can avoid mapping artefacts by creating multiply imputed maps from the posterior distribution. Under the assumptions of the model, equation (8) holds – that is, the probability that a county is highlighted is independent of its population. We illustrate with the simulated-data example above; we sample from the posterior distribution of the vector of county

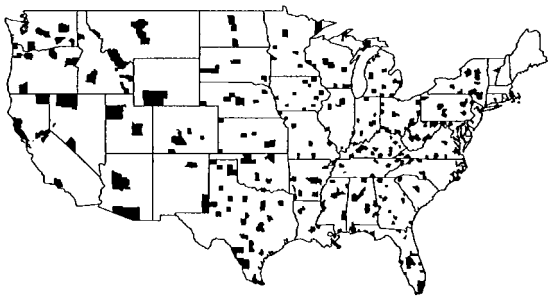


Fig. 4. Shaded counties are those in which the true county parameters θ_j are in the top 10 per cent of U.S. counties. Values of θ_j are drawn independently from a common distribution; this is thus equivalent to a selection of U.S. counties chosen at random. This map is the 'truth' that is estimated in Figures 5 and 6

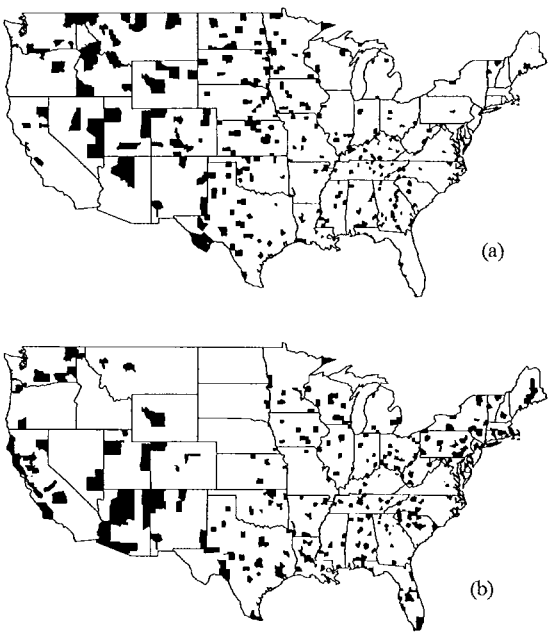


Fig. 5. (a) Shaded counties are those in which the observed rates, y_j/n_j, are in the top 10 per cent of U.S. counties. (b) Shaded counties are those in which the posterior means, $E(\theta_j|y_j) = (\alpha + y_j)/(\beta + n_j)$, are in the top 10 per cent. Compare these maps to the map of the highest true county parameters in Figure 4. The map of the observed rates highlights too many low-population rural counties, whereas the map of the posterior means includes to many high-population urban counties. These effects are perhaps most easily seen in the generally low-population counties of the Plains states

parameters – which, for the Poisson-gamma model described above, happens to be a gamma(α', β') distribution for each county with α' and β' given by the numerator and denominator of equation (9), respectively. Figure 6 displays four maps of independent multiple imputations of the vector θ, each displaying the counties with highest imputed values of θ_j. These maps differ from each other, and from Figure 4, because of the Poisson variability in the data.

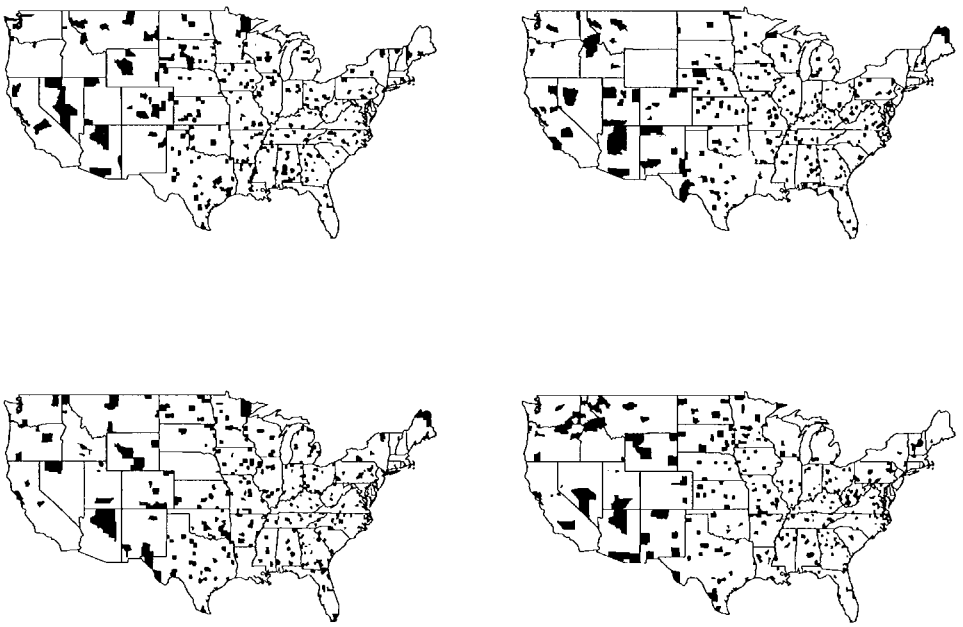


Fig. 6. Four multiple imputations. For each map, the shaded counties are those in which the imputed rates, θ_j, drawn from their posterior distribution, are in the top 10 per cent of U.S. counties, for that imputation. Compare these maps to the map of the highest true county parameters in Figure 4. These maps have no systematic artefacts due to variation in the county populations

The variation among the four maps gives some indication of the posterior uncertainty in the county parameter estimates. For example, in the map on the upper right, the western state of Wyoming has no highlighted counties, whereas in the other maps several Wyoming counties are highlighted. This implies that, given the model and the data, the true rates in those counties could be mostly low, or mostly high, or a mixture, and the maps show various of these possible realities. No strong conclusions can be drawn from any single map – in the presence of statistical uncertainties there is no way to map reality, just possible realities given the model and the data. Instead, one must look for spatial patterns that persist over most of the maps. We have no hard and fast rule for how many maps to make; in this case it seems unnecessary to display more than six or seven (or fewer than three). We suggest starting by making as many as can comfortably fit on a page while allowing sufficient resolution to discern spatial patterns if they are present.

4. MORE COMPLICATED MODELS

The basic cause of the mapping artefacts is that the posterior uncertainties in the counties are unequal, and this inequality can lead to spatial patterns in maps of point estimates. More sophisticated modelling will tend to reduce the variation of uncertainties among counties but will not, in general, equalize the uncertainties altogether. Thus the artefacts described in this paper should remain, in qualitatively similar form. Here, we briefly consider the effects of three forms of

added model sophistication: accounting for uncertainty in the hyperparameters; adding regression predictors, and spatial modelling.

Our examples would gain realism by considering the hyperparameters – (μ, σ, τ) in the normal model and (α, β) in the Poisson-gamma model – as unknown and estimated from the data rather than fixed. In general we agree with Clayton and Bernardinelli[10] that it is best to average over posterior uncertainty. This would not change the essential pattern of the figures or our main results, but it would cause the lines in Figure 3 to lose sharpness in their sawtooth pattern. The multiply imputed maps would still have no sample size artefacts.

It is standard practice to include explanatory variables (such as demographics in the analysis of cancer rates,[1] or geologic indicators in the analysis of radon levels[14]) and to explicitly model spatial correlation, typically to account for missing or poorly measured spatially-correlated covariates (see, for example, Clayton and Kaldor[18] and Mollie and Richardson[21] for Bayesian examples in disease mapping, and Cressie[22] for a general review). Unfortunately, spatial modelling does not remove the artefacts discussed in this paper, although it can sometimes reduce them by diminishing parameter uncertainties. Rather than choose specific spatial models to illustrate this point, we merely point out two extreme cases for which the presence of artefacts is readily apparent.

First, the non-spatial examples in the previous sections can be thought of as spatial models in the limit of zero spatial correlation, so if there is some *small* amount of spatial correlation, the artefacts will be nearly the same as those described above.

Second, consider an opposite extreme; suppose correlation is fairly high at small spatial scales but decreases with distance. For simplicity of exposition, suppose we are interested in a large region and that some areas around the perimeter of the region are very heavily sampled, but a large interior portion has no measurements at all. Any spatial estimation or modelling procedure we are aware of (including interpolation, splines, kriging and hierarchical Bayesian methods) will tend to generate predictions for the interior that are too smooth – subunits in the interior will have predictions that are very close to one another, since there is no information that allows them to be distinguished (see Nobre and De Macedo[23] for an example with contour maps).

Details of the artefacts in more elaborate models will obviously depend on the exact nature of the models and the data. Our point is that mapping artefacts due to spatial variation in parameter uncertainties are nearly ubiquitous, whether the mapped quantities are measured values, predictions from conventional regressions, Bayesian posterior predictions, or whatever, and whether the models are spatial or not.

5. DISCUSSION

Mapping raw data can lead to spurious spatial features. For example, regions can appear highly variable because of small sample sizes in spatial sub-units (as in the radon example) or small populations (as in the cancer example), and these apparently variable regions contain a disproportionate number of very high (or low) observed parameter values. Mapping posterior means leads to the reverse problems: areas that appear too uniform because of small sample sizes or populations. Moulton *et al.*[24] discuss some other problems with maps of posterior means. Similar problems occur with mapping counties based on statistical significance, as discussed in this article.

One way to avoid these artefacts is to produce multiple maps based on imputations from the posterior distribution (of county means, for example); spatial correlation in these maps must come from some other source. In a typical application, one might make maps of imputations from the posterior distribution of *residuals* from predictions based on covariates. Substantial spatial

correlation in the residuals that occurs in all or most of the imputed maps would indicate the presence of un-included covariates that are themselves spatially correlated, such as geologic or house construction features in the radon example. When used in this manner, multiply imputed maps can be thought of as posterior predictive checks.[25, 26]

Unfortunately, multiply imputed maps are not suitable for presenting final results (estimated cancer rates, mean radon concentrations, etc.) to most audiences, who would likely just be confused by them. Furthermore, maps really do make convenient look-up tables (what is the cancer rate, or mean radon level, in my county?). Unfortunately, even maps that are intended to be used only as look-up tables are almost sure to be used for identifying spatial features – we find it very hard to suppress this instinct ourselves. For example, a state Department of Health might map posterior estimates of county mean radon concentrations and choose to focus public education efforts on the areas of the state that appear to have high radon levels. If some contiguous group of counties is sparsely sampled – a common occurrence in practice – then these counties are likely to have near-average posterior estimated levels even if some of the counties have quite high radon levels. Therefore the group of counties will appear both average and uniform on the map, which may lead to seriously incorrect inference if the visual appearance of a large, uniform area on the map is interpreted as evidence of spatial smoothness of county mean radon levels in the area.

To the extent that some of the features identified by conventional mapping methods may be (in some cases are likely to be) artefacts, the natural tendency to associate uniformity on the map with uniformity in reality is unfortunate. Perhaps hatching or shading can be used to indicate not only the point estimates of the quantities of interest but also their uncertainties (for example, see Carlin and Louis[27]); or two maps can be presented, one of posterior means and one of posterior standard deviations; but this is a graphical design issue rather than a statistical one.

Our main goal in this paper has been to illustrate and quantify the extent to which statistical artefacts lead to misleading maps. It is clear that there are serious drawbacks to using spatial distributions of mapped point estimates to gauge the spatial distribution of quantities of interest. Multiple imputation can help avoid this problem in exploratory analysis and model checking, but we know of no satisfactory solution to the problem of generating maps for general use.

ACKNOWLEDGEMENTS

We thank Donald Rubin and Hal Stern for helpful comments. This work was supported in part by the U.S. National Science Foundation grants DMS-9404305, SBR-9708424, and Young Investigator Award DMS-9457824, and the Director, Office of Energy Research, Office of Health and Environmental Research, Environmental Services Division of the U.S. Department of Energy, under contract DE-AC03-76SF00098.

REFERENCES

1. Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P. and Pellom, A. C. 'Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates', *Journal of the American Statistical Association* **84**, 637–650 (1989).
2. Smans, M. and Esteve, J. 'Practical approaches to disease mapping', in Elliot, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, Oxford University Press, Oxford, 1992, pp. 141–150.
3. Riggan, W. B., Creason, J. P., Nelson, W. C., Manton, K. G., Woodbury, M. A., Stallard, E., Pellom, A. C. and Baubier, J. *U.S. Cancer Mortality Rates and Trends*, 1950–1979 *Vol.* IV: Maps, U.S. Government Printing Office, Washington, D.C., 1987.

4. Devine, O. J., Louis, T. A. and Halloran, M. E. 'Empirical Bayes methods for stabilizing incidence rates before mapping', *Epidemiology*, **5**, 622–630 (1994).
5. Pickle, L. W. and White, A. A. 'Effects of the choice of age-adjustment method on maps of death rates', *Statistics in Medicine*, **14**, 615–627 (1995).
6. Efron, B. and Morris, C. 'Data analysis using Stein's estimator and its generalizations', *Journal of the American Statistical Association*, **70**, 311–319 (1975).
7. Rubin, D. B. 'Using empirical Bayes techniques in the law school validity studies (with discussion)', *Journal of the American Statistical Association*, **75**, 801–827 (1980).
8. Clayton, D. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics*, **43**, 671–681 (1987).
9. Louis, T. A. 'Estimating a population of parameter values using Bayes and empirical Bayes methods', *Journal of the American Statistical Association*, **79**, 393–398 (1984).
10. Clayton, D. and Bernardinelli, L. 'Bayesian methods for mapping disease risk', in Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, Oxford University Press, Oxford, 1992, pp. 205–220.
11. Wirth, S. *et al.* 'National radon database documentation: the EPA state/residential radon surveys', Sanford Cohen and Associates. Prepared for the U.S. Environmental Protection Agency, Washington, D.C., 1992.
12. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. *Graphical Methods for Data Analysis*, Wadsworth, Pacific Grove, California, 1983.
13. Price, P. N. 'Predictions and maps of county mean indoor radon concentrations in the mid-atlantic states', *Health Physics*, **72**, 893–906 (1997).
14. Price, P. N., Nero, A. V. and Gelman, A. 'Bayesian prediction of mean indoor radon concentrations for Minnesota counties', *Health Physics*, **71**, 922–936 (1996).
15. Tufte, E. R. *The Visual Display of Scientific Information*, Graphics Press, Cheshire, Connecticut, 1983.
16. Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J. and Fraumeni, J. F. *Atlas of Cancer Mortality for U.S. Counties*: 1950–1969, Public Health Service, National Institutes of Health, Washington, D.C., 1975.
17. Schlattmann, P., Dietz, E. and Bohning, D. 'Covariate adjusted mixture models and disease mapping with the program Dismapwin', *Statistics in Medicine*, **15**, 919–929 (1996).
18. Muir, C. S. 'Cancer mapping: overview and conclusions', *Recent Results in Cancer Research*, **114**, 269–273 (1989).
19. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.
20. Rubin, D. B. 'Multiple imputation after 18+ years', *Journal of the American Statistical Association*, **91**, 473–489 (1996).
21. Mollie, A. and Richardson, S. 'Empirical Bayes estimates of cancer mortality rates using spatial models', *Statistics in Medicine*, **10**, 95–112 (1991).
22. Cressie, N. A. C. *Statistics for Spatial Data*, revised edition, Wiley, New York, 1993.
23. Nobre, F. F. and De Macedo, M. M. A. 'Feasibility of contour mapping epidemiological data with missing values', *Statistics in Medicine*, **14**, 605–613 (1995).
24. Moulton, L. H., Foxman, B., Wolfe, R. A. and Port, F. K. 'Potential pitfalls in interpreting maps of stabilized rates', *Epidemiology*, **5**, 297–301 (1994).
25. Rubin, D. B. 'Bayesianly justifiable and relevant frequency calculations for the applied statistician', *Annals of Statistics*, **12**, 1151–1172 (1984).
26. Gelman, A., Meng, X. L. and Stern, H. S. 'Posterior predictive assessment of model fitness via realized discrepancies (with discussion)', *Statistica Sinica*, **6**, 733–807 (1996).
27. Carlin, B. P. and Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London, 1996.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3235–3244 (1999)

MODELLING FOR COST-EFFECTIVENESS ANALYSIS

LOUISE B. RUSSELL*

Institute for Health, Health Care Policy, and Aging Research, Rutgers University, 30 College Avenue, New Brunswick, NJ 08901, U.S.A.

SUMMARY

A model creates the framework for a cost-effectiveness analysis, allowing decision makers to explore the implications of using an intervention in different ways and under different conditions. To serve its purpose a model must produce accurate predictions and allow for substantial variation in the factors that influence costs and effects. This paper considers three aspects of modelling: validating effectiveness estimates; modelling costs; and the implications of common statistical forms. Validation procedures similar to those for effectiveness estimates are proposed for costs. Modellers need to pay more attention to ensuring that the pathway of events described by a model represents costs as well as it does effects. Modellers can also help improve the epidemiological and clinical research on which cost-effectiveness analyses depend by showing the implications for resource allocation of the statistical forms conventionally used in these fields.

INTRODUCTION

Cost-effectiveness analysis requires estimation of the health effects and resource costs associated with an intervention and with the alternatives to which it will be compared. Modelling is frequently necessary since few studies provide information over sufficiently long periods or for all relevant costs, effects and population groups.

This paper discusses three issues faced by analysts who develop models for cost-effectiveness analysis: validating effectiveness estimates; modelling costs; and modelling functional form. Methods for validating effectiveness estimates are discussed and similar procedures are proposed for cost estimates. Modellers need to give more attention to ensuring that the pathway of events described by a model represents costs as well as it does effects. Modellers can also help decision makers by reflecting back to researchers the implications for resource allocation of the statistical forms conventionally used for health effects. At all stages of the modelling process, it is important to ensure that a model accommodates realistic differences in incidence, cost, production technology, and other conditions so that decision makers can explore the implications of providing interventions in real-world settings.

Cost-effectiveness analysis helps inform different types of decisions about health interventions. To begin, it can inform the decision to use an intervention at all by showing whether it is cost-effective enough compared to alternatives. More often decisions concern how to use the intervention. Should screening for hypertension be done every year, every two years, or every five

* Correspondence to: Louise B. Russell, Institute for Health, Health Care Policy, and Aging Research, Rutgers University, 30 College Avenue, New Brunswick, NJ 08901, U.S.A. E-mail: lrussell@rci.rutgers.edu

CCC 0277–6715/99/233235–10$17.50

years? If hypertension is diagnosed, and non-drug therapies are unsuccessful, which drugs should be used?[1] Should folic acid supplementation be accomplished through diet, vitamin supplements, or fortification of cereal grains? If fortification, how many milligrams of folic acid per 100 grams of cereal grain product?[2] Should every patient who presents at the emergency room with chest pain be monitored in the CCU or only those with other symptoms of a heart attack?[3] The best choice, moreover, can depend on how the intervention is used, the prevalence of the condition in the population, and the presence of significant risk factors or co-morbidities.

A model creates the framework for cost-effectiveness analysis. To serve its purpose, and enable decision makers to explore the implications of variation in the intervention, the condition, and the population, it must allow not only for a large number of factors to influence outcomes but also for substantial variation in those factors. In this paper I discuss three aspects of modelling for CEA: validating effectiveness estimates; modelling costs; and the implications of common statistical forms. In each case, I am particularly interested in the ability of models to simulate a realistic range of possibilities for applying an intervention.

VALIDATING EFFECTIVENESS ESTIMATES

Accuracy is essential for a model. Does it accurately predict the outcomes that would occur if an intervention is applied? A decade ago, Eddy described four levels of validation.[4] First, the structure of the model should make sense to experts. Second, the model should reproduce the outcomes observed in the studies used to estimate its parameters. Third, the model's predictions could be compared with results from studies not used in its construction. Fourth, the model could be used to predict outcomes for a new programme and the predictions compared with the outcomes when the programme is implemented. The first and second steps are essential. For the third step, randomized clinical trials (RCTs) offer a challenging, but potentially persuasive, test of a model's accuracy.

Two groups of investigators have tested models of coronary heart disease (CHD) against trial results. Grover *et al.*[5] constructed a model using risk relationships from Framingham and data for the Canadian population. They projected outcomes for cohorts with the same average risk factors as those in three RCTs: MRFIT; the Lipid Research Clinics Trial; and the Helsinki Heart Study. The predictions of CHD events and deaths matched well, almost always falling within the 95 per cent confidence interval for the corresponding trial result. For the MRFIT and Helsinki trials, CHD event rates could be compared year by year, and predicted rates fell in the 95 per cent confidence interval in 25 of 26 cases.

Hlatky *et al.*[6] developed a model to predict survival based on Cox proportional hazard regressions fitted to clinical and laboratory measures in the Duke Cardiovascular Disease Databank. For their comparisons, they selected patients from the Duke database who would have qualified for the trial in question. They used the model to predict survival at five years for those patients, with medication and with surgery, and compared the predictions with the 95 per cent confidence intervals for the corresponding trial groups.

A total of 719 Duke patients would have qualified for the Veterans Administration Cooperative Study. With one exception, the predictions fell within the confidence intervals, for all patients and by number of diseased vessels. Overall, five-year survival in the medical group was 78 per cent in the VA trial (95 per cent CI, 73–82); the model predicted 80·9 per cent. It was 83 per cent in the surgical group (95 per cent CI, 79–87); the model predicted 85·5 per cent.

In the VA trial survival was lower for surgical patients with two-vessel disease than those with three-vessel disease. This finding disagreed with the model predictions and with 'a large body of data showing that patients with two vessel disease have a better prognosis'. The authors suggested that the trial result might be due to the substantial sampling variability that can affect results for small subgroups. The example suggests that comparisons are not always a one-way street from the correct trial to the less accurate model. While trials are usually the benchmark, the model may be more accurate on specific points.

The authors also made projections for 512 Duke patients who would have qualified for the European Cooperative Surgery Study and 250 who would have qualified for the Coronary Artery Surgery Study (CASS). Again, the predictions fell within the 95 per cent confidence intervals for the corresponding trial endpoints. The model, however, predicted substantially lower survival for the medically treated participants in CASS than actually occurred. In that trial, the sickest patients in the group initially randomized to medication were withdrawn and sent to surgery. The authors hypothesized that, had assignment remained random, the trial results might have been closer to the predictions.

An obvious requirement for comparisons, not always easy to meet using published trial reports, is that predictions be made for the kinds of patients enrolled in the trial. The best model cannot be expected to predict outcomes similar to those of a trial unless the patient characteristics used for the predictions are approximately the same as those of the trial participants. Interestingly, Hlatky *et al.* were able to suggest a reconciliation of results in two trials based on differences in their participants. In the European trial, patients with three vessel disease treated surgically had a substantially higher survival rate than those treated medically. No such advantage appeared in CASS. After comparing their predictions with trial results, by number of diseased vessels and ejection fraction, the authors suggested that those randomized to surgery in CASS were at lower risk than those in the European trial. They concluded (reference 6, p. 244):

> 'If both trials are accepted as correct in their patient populations, the results suggest that high risk patients with three vessel disease receive a survival benefit from surgery, whereas low risk patients … do not appear to benefit'.

In these studies, models compared well with trials available at the time the models were constructed, but not used in their construction. Models have also performed well when compared with trial results published after the model was built. To mention two cases, Weinstein and Stason's model of mild hypertension was based on statistically insignificant differences in a VA trial;[7] these were later confirmed by the large Australian trial and other trials.[8] On the basis of early studies, Siegel *et al.* analysed the use of angiotensin-converting enzyme inhibitors to retard kidney damage in patients with insulin-dependent diabetes;[9] their effectiveness assumptions were later confirmed by an RCT.[10]

It is reasonable to expect a good model to match the results of trials available at the time of its construction, but not to expect it to predict the results of future trials. Models can and should accurately reflect the state of knowledge at the time they are created. If that knowledge were definitive, trials would not be ethical. When it is not definitive, trials will sometimes fail to confirm beliefs that were widely held before they were completed. At that point, both medical practice and models need to be revised.

None the less, a common view is that models are doubtful vehicles, full of 'assumptions', a word which seems to call up in people's mind's adjectives like 'untested', 'untrustworthy' and

'spurious'. Brown and Fintor addressed this perception in a review of studies of breast cancer screening for women 50 years and older, which reported widely different cost-effectiveness ratios.[11] They compared two of the studies in detail: a U.S. study published in 1987, and a Dutch study published in 1991. When the two models were used to calculate cost-effectiveness ratios for the same set of assumptions, the results were almost identical: $7256 per life-year gained in the U.S.; $7250 in the Netherlands. The differences in the published results were caused by different assumptions, not by some mysterious process buried deep in the model itself.

The original studies made different assumptions because they were analysing different problems. Suspicion of models may arise as much from the common reaction to the word 'assumption' as from differences in results. Consider the following statements.

(i) The models produced different results because they used different assumptions.
(ii) The models produced different results because, although they were dealing with the same disease and the same intervention, they analysed different ways of applying it in different populations.

The first statement can make it sound as though differences should not exist. The second makes it obvious that they should and will. The U.S. and Dutch studies produced different results largely because the U.S. study analysed annual screening with aggressive follow-up at U.S. prices. To get the same results from both models the authors had to make 'the contrived assumption that "Dutch" conditions for screening mammography – low priced biennial screening and conservative follow-up procedures – prevail(ed) in the United States'. There were also differences in the costs included and length of follow-up, differences that the recommendations of the Panel on Cost-Effectiveness in Health and Medicine seek to reduce.[12] However, the larger part arose from the different circumstances the studies analysed.

A major purpose of models is to allow analysts and decision makers to explore the effectiveness and cost-effectiveness of interventions under different circumstances. Exploration can be particularly useful when direct evidence is scant.[13,14] When this means exploring the implications of a higher or lower cost for the intervention, most people feel comfortable going beyond the evidence. When it means exploring the implications of applying the intervention to a group that was not included in trials or epidemiological studies, the level of comfort is lower.

When is a model going too far beyond the data? The usual answer is couched in terms of the quality of the data. I would propose instead that medical and public health practice are the best guides. Models can appropriately be used to analyse any circumstances in which the intervention is already being applied, or in which it is being seriously considered for application. If it is appropriate to use the intervention in the real world, on real people, it is appropriate to analyse the implications of that use with a model.

Brown and Fintor remind us that, because the circumstances in which an intervention can be applied are so varied, the test of a model's accuracy is not whether it produces the same results for the problem it was designed to analyse as another model produces for another problem. Effectiveness and cost-effectiveness are not inherent, unchanging properties of an intervention. They can differ substantially with the characteristics of the population, the way the intervention is applied, and its costs. When models produce very different results for the same problem, there is cause for concern, but not when they produce different results for different problems.

MODELLING COSTS

Modelling has emphasized events that influence an intervention's effectiveness. Milestones in the progress of a condition and treatment decisions make up the pathway of events for which modellers collect data. Costs are assigned to these events as appropriate, but it is rare for the events to be thought through in the first place with as much attention to costs as to effects. The assumption seems to be that costs follow effects and that if effects are well represented, costs will be taken care of automatically.

It is time to extend our notions of model validation to the validation of costs. Consider Eddy's four suggestions. The first was that the structure of the model should make sense to experts. Modellers could seek out experts on resource use and costs to review the model for accuracy and completeness. If this is done currently, effectiveness experts are asked to review the model for both effects and costs, despite evidence that doctors are often only vaguely aware of the costs of what they do.

The second level of validation was that models should reproduce the data used in their creation. If Medicare was the source of the data on individual episodes of resource use and costs, does the model reproduce Medicare's expenditures for the entire pathway of events? If resource use and cost per item came from a university hospital, does the model reproduce the hospital's total expenditure for the course of care? Questions like these are rarely asked or answered. Yet it is as just as important for cost-effectiveness analyses to predict costs accurately as to predict effects accurately.

The third level was that models could be tested against data not used in their construction. Attempts are under way to collect cost data in clinical trials. For these trials, models could be tested for their ability to predict trial costs as well as health outcomes. A recent article on the cost-effectiveness of regimens to treat *Pneumocystis carinii* pneumonia in AIDS patients noted that it would be desirable to run such tests.[15] The authors updated the effectiveness parameters in their model to match data from an RCT, and compared the cost-effectiveness ratios with those they had reported earlier, but they could not update their cost information because the trial had not collected cost data.

In part, the failure to validate cost estimates reflects the failure to take cost data as seriously as effectiveness data. Effects are usually estimated on the basis of a careful review of epidemiological and clinical trial data. Costs are often gathered from the nearest convenient source. Effects may apply to the U.S. population, costs to Philadelphia or Boston. A basic requirement for accurate predictions, often overlooked, is that both costs and effects should apply to the same population and the same circumstances. Further, data on resource use and costs need to be collected with the same care and subjected to the same sorts of consistency checks as effectiveness data – comparing one source with another, relating differences in costs to characteristics thought to be associated with those differences, and so on. Wolff and colleagues have documented the substantial errors that can occur when cost data are taken uncritically from a single source.[16,17]

In addition, the range of variation that could usefully be modelled is as wide for costs as for effects. An intervention's effectiveness differs across the country because populations differ in incidence of the condition, risk factors and co-morbidities. Costs differ across the country because of differences in wages and other costs, in practice patterns and in suitable production technologies. While studies usually report sensitivity analyses on major costs individually, they rarely model costs in different parts of the country. They report the results when a drug costs half or twice the baseline value, but not the results when the intervention is delivered in South Carolina

or Alaska instead of New York, or in an HMO instead of a fee-for-service practice. While one purpose of sensitivity analyses is to determine which parameters have a major influence on cost-effectiveness, it would also be useful to explore sets of assumptions that describe, as accurately as the data allow, circumstances in another part of the country or another delivery system.

Some studies have examined how cost-effectiveness would be influenced by different production techniques. Eisenberg *et al.*[18] compared a cephalosporin antibiotic that could be given once a day with conventional cephalosporins that had to be given three or four times a day. Using time-and-motion studies, they documented differences in labour and materials associated with the two schedules and estimated national savings from the once-a-day regimen at $85 million to $115 million per year.

That difference would not have appeared in most analyses because resource use could not have been measured at that level of detail from the usual sources. The authors had to do the equivalent for costs of an observational study of health outcomes. The Panel on Cost-Effectiveness in Health and Medicine has urged the use of micro-costing for costing events important to an analysis.[12] Micro-costing could yield a better understanding of the factors that underlie resource use and costs for various conditions, analogous to the understanding of effectiveness built up from epidemiological and clinical research. That understanding might reveal alternatives for making interventions more cost-effective by changing the way they are delivered, not just by targeting them to population subgroups.

Schechtman *et al.* examined two different production techniques for managing elevated cholesterol.[19] In an RCT, 247 veterans with elevated cholesterol were assigned to usual care or an experimental programme (physician extenders trained to start with the most cost-effective therapies). Over two years, the experimental programme was found to be more expensive, but also more effective, so that its cost per unit reduction in LDL cholesterol was lower. This study is a pioneering example of the use of an RCT to evaluate both sides of the cost-effectiveness equation.

Since Eddy's classic study of screening for cervical cancer,[20] it has become commonplace to evaluate different screening frequencies, which have a major influence on costs. Some authors have compared different dosages of drugs, for example the study of cholesterol-reducing drugs by Goldman *et al.*[21] Analysing hypothetical, but plausible, production alternatives can also suggest possibilities for making an intervention more cost-effective. For example, Schulman *et al.* examined the impact of developing a test that could quickly identify patients with Gram-negative sepsis on the cost-effectiveness of Centoxin.[22,23]

Models should be flexible enough to permit exploration of a range of production possibilities and cost levels for an intervention. Analysts could then examine plausible differences in costs and production technologies. It would be useful to evaluate combinations of values that occur in the real world: conditions in Michigan versus those in San Francisco, conditions in an inner city, a suburb, or a rural area.

MODELLING FORM

Eddy has written that the most common error in model-building is oversimplification and that '(t)he most common causes of oversimplification are to omit important variables and to attempt to squeeze a problem into a familiar or convenient mathematical form, rather than to create a form to fit the problem'.[4]

Table I. Cost-effectiveness of reducing serum cholesterol

Age (years)	Initial cholesterol (mg/dl)	Cost per year of life gained*	
		Low-risk patient†	High-risk patient‡
40	240	$180,000	$21,000
	300	94,000	11,000
60	240	280,000	23,000
	300	160,000	13,000

* Discount rate is 5 per cent
† Low risk defined as no cigarette smoking, systolic blood pressure at the 10th percentile of the age- and sex-specific population distribution, and high-density lipoprotein cholesterol at the 90th percentile of the age- and sex-specific population distribution.
‡ High risk defined as cigarette smoking habit, systolic blood pressure at the 90th percentile of the age- and sex-specific population distribution, and high-density lipoprotein cholesterol at the 10th percentile of the age- and sex-specific population distribution
Source: reference 24, Table 44.3, p. 440

Models are built from estimates of risk – the probability that a condition will progress to the next stage, that a test is accurate, that a treatment will be effective. In medical research, the familiar and convenient mathematical forms for fitting risk relationships are the logistic, and, more recently, hazard models. Both forms incorporate an assumption that the risk relationship is multiplicative, and thus that the size of the risk reduction caused by changing one risk factor differs for different levels of the other risk factors. This assumption implies, for example, that the reduction in risk caused by lowering systolic blood pressure from 160 mmHg to 140 mmHg will be larger in people who also smoke, even though they continue to smoke, than in people whose only risk factor is high blood pressure. Similarly, the reduction in risk from smoking cessation will be greater in people who are hypertensive, even if their blood pressure is unchanged, than in non-smokers.

In turn, this implies that it will be more cost-effective to apply an intervention to people with several risk factors, not because the programme achieves economies by treating several risk factors, but because intervening against a single risk factor is more effective in these people. The point is clear in an analysis by Taylor *et al.* of a dietary programme to lower serum cholesterol modelled after the one employed in MRFIT.[24] Effectiveness was estimated using logistic coefficients reported from the Framingham Study. Results were presented separately for low-risk men, whose only risk factors for heart disease were their gender and cholesterol level, and for high-risk men, who also smoked and had high blood pressure and low HDL levels. Although the cost of the intervention was the same, cost per life-year was approximately ten times higher for low-risk men because of the multiplicative assumption incorporated in the logistic form (Table I).

The assumption is an inherent property of these forms and has not been tested until recently. Psaty *et al.* found that the apparent diminution with age in the strength of some risk factors for heart disease might be an artefact of the conventional models.[25] Similarly, Silberberg found that

the relationships between CHD death and serum cholesterol, smoking, and hypertension were 'closer to additive than multiplicative'.[26] Russell *et al.* found that an additive model predicted heart disease mortality for participants in the NHANES I Epidemiologic Followup Study as accurately as the logistic.[27]

Logistic and hazard models play an important role in some of the situations for which models are particularly useful – examining differences in effectiveness and cost-effectiveness among subgroups. When analysts model the implications of targeting an intervention to subgroups, or extrapolate to explore its application to less-studied groups, they need to be aware of the implications of the conventional forms. Modellers cannot supply the data to resolve this issue, but they can draw attention to it by showing how estimates change when additive and multiplicative forms are used. The ultimate goal is to ensure that estimated differences among subgroups are not an artefact of a convenient statistical model.

CONCLUSIONS

In this paper I have reviewed validation procedures used for effectiveness estimates and have proposed that the same kinds of procedures be adapted for application to cost estimates. More generally, modellers need to pay more attention to ensuring that the pathway of events described by a model represents costs as well as it does effects and that the model is flexible enough to incorporate the range of costs and production technologies faced by decision makers. Modellers can also help decision makers by reflecting back to clinical and epidemiological researchers the implications for resource allocation of conventional statistical forms.

The purpose of cost-effectiveness analysis is to compare alternative ways of investing in health. Analyses should minimize unnecessary differences in order to ensure that the comparisons are as accurate as possible. The recommendations of the Panel on Cost-Effectiveness in Health and Medicine were directed to this goal.[12] At the same time, models need to be constructed so that they can mirror real-world conditions. Analyses of the same intervention applied differently to different populations will produce different cost-effectiveness ratios. This is no reason to doubt models. A major part of their usefulness is that they allow decision makers to explore the value of investing in an intervention under different conditions before resources are committed.

ACKNOWLEDGEMENTS

I am grateful to Dennis Fryback, Alan Garber, Joanna Siegel, Frank Sonnenberg and Milton Weinstein for suggesting CEA studies to illustrate the points made in this paper and to Dennis Fryback for sending me several papers on modelling, including the pages by David Eddy from the IOM report *Assessing Medical Technologies.*

REFERENCES

1. Edelson, J. T., Weinstein, M. C., Tosteson, A. N. A. *et al.* 'Long-term cost-effectiveness of various initial monotherapies for mild to moderate hypertension', *Journal of the American Medical Association*, **263**, 407–413 (1990).
2. Kelly, A. E., Haddix, A. C., Scanlon, K. S. *et al.* 'Cost-effectiveness of strategies to prevent neural tube defects', in Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein M. C. (eds), *Cost-Effectiveness in Health and Medicine*, Oxford University Press, New York, 1996.
3. Fineberg, H. V., Scadden, D. and Goldman, L. 'Care of patients with a low probability of acute myocardial infarction', *New England Journal of Medicine*, **310**, 1301–1307 (1984).

4. Eddy, D. M. 'Technology assessment: The role of mathematical modeling', in Committee for Evaluating Medical Technologies in Clinical Use, Institute of Medicine, *Assessing Medical Technologies*, National Academy Press, Washington, D.C., 1985, 144–154.
5. Grover, S. A., Abrahamowicz, M., Joseph, L. *et al.* 'Benefits of treating hyperlipidemia to prevent coronary heart disease: estimating changes in life expectancy and morbidity', *Journal of the American Medical Association*, **267**, 816–822 (1992).
6. Hlatky, M. A., Califf, R. M., Harrell, F. E. *et al.* 'Comparison of predictions based on observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery', *Journal of the American College of Cardiology*, **11**, 237–245 (1988).
7. Weinstein, M. C. and Stason, W. B. 'Economic considerations in the management of mild hypertension', *Annals of the New York Academy of Sciences*, **304**, 424–440 (1978).
8. Management Committee. 'The Australian therapeutic trial in mild hypertension', *Lancet*, **1**(8181), 1261–1267 (1980).
9. Siegel, J. E., Krolewski, A. S., Warram, J. H. and Weinstein, M. C. 'Cost-effectiveness of screening and early treatment of nephropathy in patients with insulin-dependent diabetes mellitus', *Journal of the American Society of Nephrology*, **3**, S111–S119 (1992).
10. Lewis, E. J., Hunsicker, L. G., Bain, R. P. *et al.* 'Effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy', *New England Journal of Medicine*, **329**, 1456–1462 (1993).
11. Brown, M. L. and Fintor, L. 'Cost-effectiveness of breast cancer screening: preliminary results of a systematic review of the literature', *Breast Cancer Research and Treatment*, **25**, 113–118 (1993).
12. Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. *Cost-Effectiveness in Health and Medicine*, Oxford University Press, New York, 1996.
13. Fryback, D. G. and Frame, P. S. 'Screening for abdominal aortic aneurysms: time-based modeling for public policy', in Swain, J. J., Goldsman, D., Crain, R. C. and Wilson, J. R. (eds), 1992 *Winter Simulation Conference Proceedings*, meetings held at Arlington, VA, December 13–16, 1992.
14. Siegel, J. E., Weinstein, M. C. and Fineberg, H. V. 'Bleach programs for preventing AIDS among IV drug users: modeling the impact of HIV prevalence', *American Journal of Public Health*, **81**, 1273–1279 (1991).
15. Freedberg, K. A., Hardy, W. D., Holzman, R. S. *et al.* 'Validating literature-based models with direct clinical trial results: the cost-effectiveness of secondary prophylaxis for PCP in AIDS patients', *Medical Decision Making*, **16**, 29–35 (1996).
16. Wolff, N. and Helminiak, T. W. 'Nonsampling measurement error in administrative data: Implications for economic evaluations', *Health Economics*, **5**, 501–512 (1996).
17. Wolff, N., Helminiak, T. W. and Diamond, R. J. 'Estimated societal costs of assertive community mental health care', *Psychiatric Services*, **46**, 898–906 (1995).
18. Eisenberg, J. M., Koffer, H. and Finkler, S. A. 'Economic analysis of a new drug: potential savings in hospital operating costs from the use of a once-daily regimen of parenteral cephalosporin', *Reviews of Infectious Diseases*, **6**, S909–S923 (1984).
19. Schectman, G., Wolff, N., Byrd, J. C., Hiatt, J. G. and Hartz, A. 'Physician extenders for cost-effective management of hypercholesterolemia', *Journal of General Internal Medicine*, **11**, 277–286 (1996).
20. Eddy, D. M. *Screening for Cancer*: *Theory*, *Analysis*, *and Design*, Prentice-Hall, 1980.
21. Goldman, L., Weinstein, M. C., Goldman, P. A. and Williams, L. W. 'Cost-effectiveness of HMG-CoA reductase inhibition for primary and secondary prevention of coronary heart disease', *Journal of the American Medical Association*, **265**, 1145–1151 (1991).
22. Schulman, K. A., Glick, H. A., Rubin, H. and Eisenberg, J. M. 'Cost-effectiveness of HA-1A monoclonal antibody for gram-negative sepsis', *Journal of the American Medical Association*, **266**, 3466–3471 (1991).
23. Fisher, L. M. 'Investors punish Centocor for more bad news', *New York Times*, 19 January, 1993, C1.
24. Taylor, W. C., Pass, T. M., Shepard, D. S. and Komaroff, A. L. 'Cost-effectiveness of cholesterol reduction for the primary prevention of coronary heart disease in men', in Goldbloom, R. G. and Lawrence, R. S. (eds), *Preventing Disease*: *Beyond the Rhetoric*, Springer-Verlag, New York, 1990, pp. 437–441.

25. Psaty, B. M., Koepsell, T. D., Manolio, T. A. *et al*. 'Risk ratios and risk differences in estimating the effect of risk factors for cardiovascular disease in the elderly', *Journal of Clinical Epidemiology*, **43**, 961–970 (1990).
26. Silberberg, J. S. 'Estimating the benefits of cholesterol lowering: are risk factors for coronary heart disease multiplicative?', *Journal of Clinical Epidemiology*, **43**, 875–879 (1990).
27. Russell, L. B., Taylor, W. C., Jagannathan, R. *et al*. 'What statistical model best describes heart disease risk? Evidence from the NHANES I Epidemiologic Followup Study', Institute for Health, Health Care Policy, and Aging Research, Rutgers University, New Brunswick, NJ, working paper, 1996.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3245–3262 (1999)

CONSTRUCTING CONFIDENCE INTERVALS FOR COST-EFFECTIVENESS RATIOS: AN EVALUATION OF PARAMETRIC AND NON-PARAMETRIC TECHNIQUES USING MONTE CARLO SIMULATION

ANDREW H. BRIGGS[1]*, CHRISTOPHER Z. MOONEY[2] AND DAVID E. WONDERLING[3]

[1] *Health Economics Research Centre, Oxford Institute of Health Sciences and Nuffield College, University of Oxford, U.K.*
[2] *Department of Political Science, West Virginia University, U.S.A*
[3] *Health Promotion Sciences Unit, London School of Hygiene and Tropical Medicine, U.K.*

SUMMARY

The statistic of interest in most health economic evaluations is the incremental cost-effectiveness ratio. Since the variance of a ratio estimator is intractable, the health economics literature has suggested a number of alternative approaches to estimating confidence intervals for the cost-effectiveness ratio. In this paper, Monte Carlo simulation techniques are employed to address the question of which of the proposed methods is most appropriate. By repeatedly sampling from a known distribution and applying the different methods of confidence interval estimation, it is possible to calculate the coverage properties of each method to see if these correspond to the chosen confidence level. As the results of a single Monte Carlo experiment would be valid only for that particular set of circumstances, a series of experiments was conducted in order to examine the performance of the different methods under a variety of conditions relating to the sample size, the coefficient of variation of the numerator and denominator of the ratio, and the covariance between costs and effects in the underlying data. Response surface analysis was used to analyse the results and substantial differences between the different methods of confidence interval estimation were identified. The methods, both parametric and non-parametric, which assume a normal sampling distribution performed poorly, as did the approach based on simply combining the separate intervals on costs and effects. The choice of method for confidence interval estimation can lead to large differences in the estimated confidence limits for cost-effectiveness ratios. The importance of such differences is an empirical question and will depend to a large extent on the role of hypothesis testing in economic appraisal. However, where it is suspected that the sampling distribution is skewed, normal approximation methods produce particularly poor results and should be avoided. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

The purpose of the economic appraisal of health care interventions is to inform public health decision makers of the relative value for money (or cost-effectiveness) of funding alternative interventions. Where the intervention in question generates improved health outcomes for

* Correspondence to: Andrew Briggs, Health Economics Research Centre, Institute of Health Sciences, University of Oxford, Headington, Oxford OX3 7LF, U.K. E-mail: andrew.briggs@ihs.ox.ac.uk

Contract/grant sponsor: U.K. Department of Health
Contract/grant sponsor: Nuffield College Goodhart fund
Contract/grant sponsor: Office of Health Economics

CCC 0277–6715/99/233245–18$17.50

patients, but at increased overall cost, the appropriate summary measure of cost-effectiveness is the incremental cost-effectiveness ratio (ICER). The ICER measures the additional cost of one intervention over another (say treatment A and treatment B) per unit difference in effectiveness. Where data have been obtained from two samples of patients receiving the different treatments, the ICER is calculated as in equation (1):

$$\hat{R} = \frac{\bar{C}_A - \bar{C}_B}{\bar{E}_A - \bar{E}_B} = \frac{\Delta\bar{C}}{\Delta\bar{E}} \tag{1}$$

where $\bar{C}_A$ and $\bar{E}_A$ are the mean costs and effects for the sample receiving treatment A and $\bar{C}_B$ and $\bar{E}_B$ are the mean costs and effects for the sample receiving treatment B.

When sample data on costs and effects are available it is natural to consider the use of statistical techniques to calculate confidence intervals around such point estimates. Unfortunately, the calculation of confidence intervals for a ratio is far from straightforward since the probability of obtaining a zero or near zero value on the denominator of the ratio is non-negligible, which suggests that the moments of the ICER may be undefined. In practice, this is a very real problem since it is common for clinical trials to be designed to detect the smallest meaningful clinical difference between treatments and is likely to lead to a large number of studies showing differences in treatment effects which are close to zero.[1] Clearly, this presents a problem for the use of standard parametric statistical methods. Recent research has focused on parametric approximations to the confidence interval for the ICER.[2–5] In addition, several commentators have also proposed the non-parametric approach of bootstrapping as a method for estimating confidence intervals,[1,2,6,7] and this approach has been successfully demonstrated using clinical trial data.[5,8] In the face of all the possible methods, one question quickly surfaces. Which of these methods is the most appropriate?

In this paper we present the results of a Monte Carlo simulation exercise designed to evaluate the alternative methods of calculating confidence intervals for the ICER statistic, under a variety of different conditions. The experiments require a massive number of iterations which, even a few years ago, would have put this exercise beyond our reach, but which is now possible thanks to the increasing power of personal computers.

2. METHODS

This section is split into three. The first two parts describe in some detail the alternative parametric and non-parametric methods that have been proposed for estimating the confidence limits for the ICER. The third part of this section presents the overall Monte Carlo simulation experiment designed to evaluate each of the methods.

2.1. Parametric approaches to estimating the ICER confidence interval

Three main methods, based on the parametric approach for calculating confidence intervals for an ICER, have recently appeared in the literature.[2–5] Each of these methods is explored in turn, highlighting the assumptions on which it is based.

2.1.1. The confidence box approach

A number of commentators have advocated the cost-effectiveness plane (CE plane) for presenting the results of economic evaluation and for aiding policy decisions.[9,10] O'Brien and colleagues

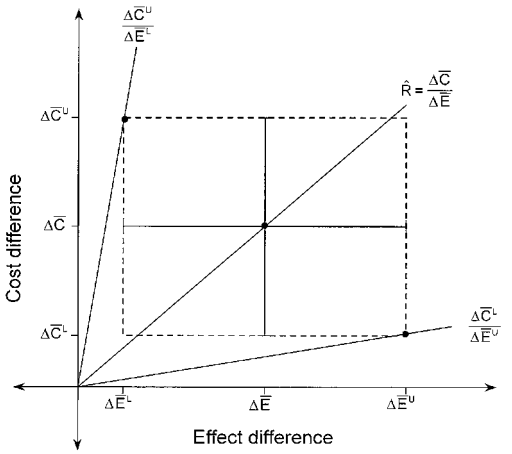


Figure 1. Confidence limits on the cost-effectiveness plane and the 'confidence box' approach to estimating confidence limits for the ICER

showed how the CE plane could also be used to present the confidence limits for the estimate of incremental cost-effectiveness.[2] Figure 1, which is based on the representation by O'Brien and colleagues, shows the results of a hypothetical prospective economic evaluation on the CE plane. The difference in effect between two therapies is shown on the horizontal axis with mean effect difference $\Delta\bar{E}$ and upper and lower confidence limits for the effect difference ($\Delta\bar{E}^{U}$ and $\Delta\bar{E}^{L}$) represented by the horizontal 'I' bar. Similarly, the difference in cost between two therapies is shown on the vertical axis with mean cost difference $\Delta\bar{C}$, and upper and lower confidence limits for the cost difference ($\Delta\bar{C}^{U}$ and $\Delta\bar{C}^{L}$) represented by the vertical 'I' bar. These 'I' bars intersect at point ($\Delta\bar{E}$, $\Delta\bar{C}$), hence the ray that connects this point of intersection to the origin has a slope equal to the value of the ICER. O'Brien and colleagues argue that combining the limits of the confidence intervals for costs and effects separately gives natural best and worst case limits on the ratio; that is, the upper limit of the cost difference over the lower limit of the effect difference ($\Delta\bar{C}^{U}/\Delta\bar{E}^{L}$) gives the highest values of the ratio (worst case) and the lower limit of costs divided by the upper limit of effects ($\Delta\bar{C}^{L}/\Delta\bar{E}^{U}$) gives the lowest (best) value of the ratio.

2.1.2. The Taylor series approximation

Rather than use these extreme limits, which are likely to overestimate the true interval, O'Brien and colleagues argue that it is possible to use the Taylor series approximation of the variance of a function of two random variables to estimate the variance of a ratio.[2] The advantage of this method is that it accounts for the covariance between the numerator and denominator. Having approximated the variance of the ICER statistic in this way, assuming the sampling distribution of the ICER to be normal allows the confidence interval to be estimated in the traditional manner.

The Taylor approximation shows that where y is a function of two random variables x_1 and x_2, the variance of y can be expressed in terms of the partial derivatives of y with respect to x_1 and x_2, weighted by the variances and covariance of x_1 and x_2. The Taylor series formula is

$$\mathrm{var}(y) \approx \left(\frac{\partial y}{\partial x_1}\right)^2 \mathrm{var}(x_1) + \left(\frac{\partial y}{\partial x_2}\right)^2 \mathrm{var}(x_2) + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right)\mathrm{cov}(x_1, x_2). \quad (2)$$

Equation (2) can now be solved for the case of the ICER presented in equation (1) by substituting $\Delta\bar{C}$ for x_1 and $\Delta\bar{E}$ for x_2.* Hence equation (3) gives the Taylor series approximation of the variance of the ratio estimator, using the sample estimates of the means and variances (since by definition, the population values cannot be observed):

$$\mathrm{var}(\hat{R}) \approx \frac{1}{\Delta\bar{E}^2}\,\mathrm{var}(\Delta\bar{C}) + \frac{\Delta\bar{C}^2}{\Delta\bar{E}^4}\,\mathrm{var}(\Delta\bar{E}) - 2\,\frac{\Delta\bar{C}}{\Delta\bar{E}^3}\,\mathrm{cov}(\Delta\bar{C}, \Delta\bar{E}). \tag{3}$$

Factoring $\hat{R}^2 = \Delta\bar{C}^2/\Delta\bar{E}^2$ from the right-hand side simplifies (3) to

$$\mathrm{var}(\hat{R}) \approx \hat{R}^2\left[\frac{\mathrm{var}(\Delta\bar{C})}{\Delta\bar{C}^2} + \frac{\mathrm{var}(\Delta\bar{E})}{\Delta\bar{E}^2} - 2\,\frac{\mathrm{cov}(\Delta\bar{C}, \Delta\bar{E})}{\Delta\bar{C}\Delta\bar{E}}\right].$$

Noting that the coefficient of variation for a random variable x is defined $\mathrm{cv}(x) = \sqrt{\mathrm{var}(x)}/\bar{x}$ and that the correlation coefficient between two random variables x and y is defined $\rho_{xy} = \mathrm{cov}(x, y)/\sqrt{\{\mathrm{var}(x)\,\mathrm{var}(y)\}}$ further simplifies the exposition:

$$\mathrm{var}(\hat{R}) \approx \hat{R}^2[\mathrm{cv}(\Delta\bar{C})^2 + \mathrm{cv}(\Delta\bar{E})^2 - 2\rho\mathrm{cv}(\Delta\bar{C})\mathrm{cv}(\Delta\bar{E})]. \tag{4}$$

Employing standard parametric assumptions gives the confidence interval as

$$(\hat{R} - z_{\alpha/2}\sqrt{\mathrm{var}(\hat{R})}, \quad \hat{R} + z_{\alpha/2}\sqrt{\mathrm{var}(\hat{R})})$$

where $z_{\alpha/2} = \Phi^{-1}[1 - \alpha/2]$, Φ^{-1} is the inverse of the cumulative distribution of the standard normal function, and $100(1 - \alpha)$ per cent is the confidence level.

O'Brien and colleagues recognize that although the assumption of a normal distribution may be justified in the case of large samples, it is unlikely that the distribution of a ratio will follow a well-behaved distribution in general.[2] However, even if samples are large, the distribution is likely to be skewed where the coefficient of variation of the denominator of the ICER (effect difference) is high.[5]

2.1.3. Fieller's method

An alternative method of calculating confidence intervals around ratios has been described by Fieller.[11] This approach has been advocated for use in calculating confidence intervals around ICERs by both Willan and O'Brien[3] and Chaudhary and Stearns.[5] The method is described in general terms by Cochran.[12]

The advantage of Fieller's method over the Taylor series expansion is that it takes into account the skew of the ratio estimator. The method assumes that the numerator and denominator of the ratio follow a joint normal distribution function such that (in the case of the ICER) $\Delta\bar{C} - R\Delta\bar{E}$ is normally distributed. Hence, dividing through by the standard deviation equation (5) follows the standard normal distribution:

$$\frac{\Delta\bar{C} - R\Delta\bar{E}}{\sqrt{\{\mathrm{var}(\Delta\bar{C}) + R^2\,\mathrm{var}(\Delta\bar{E}) - 2R\,\mathrm{cov}(\Delta\bar{C}, \Delta\bar{E})\}}} \sim \mathrm{N}(0, 1). \tag{5}$$

* The partial derivatives of the ICER with respect to $\Delta\bar{C}$ and $\Delta\bar{E}$ are $1/\Delta\bar{E}$ $-$and $\Delta\bar{C}/\Delta\bar{E}^2$, respectively.

Setting this expression equal to $z_{\alpha/2}$ and rearranging gives the following quadratic equation in R (using the simplified notation introduced in equation (4)):

$$R^2[1 - z^2_{\alpha/2}\mathrm{cv}(\Delta\bar{E})^2] - 2R\hat{R}[1 - z^2_{\alpha/2}\rho cv(\Delta\bar{E})cv(\Delta\bar{C})] + \hat{R}^2[1 - z^2_{\alpha/2}cv(\Delta\bar{C})] = 0 \qquad (6)$$

where $\hat{R}$ is defined from equation (1).

Solving equation (6) for R using the standard quadratic formula* gives the confidence interval as

$$\hat{R}\left[\frac{1 - z^2_{\alpha/2}\,\rho\mathrm{cv}(\Delta\bar{C})\mathrm{cv}(\Delta\bar{E})}{1 - z^2_{\alpha/2}\mathrm{cv}(\Delta\bar{E})^2}\right]$$
$$\pm z_{\alpha/2}\hat{R}\left[\frac{\sqrt{\{\mathrm{cv}(\Delta\bar{C})^2 + \mathrm{cv}(\Delta\bar{E})^2 - 2\rho(\Delta\bar{C})\mathrm{cv}(\Delta\bar{E}) - z^2_{\alpha/2}(\mathrm{cv}(\Delta\bar{C})^2\mathrm{cv}(\Delta\bar{E})^2 - \rho^2\mathrm{cv}(\Delta\bar{C})^2\mathrm{cv}(\Delta\bar{E})^2)\}}}{1 - z^2_{\alpha/2}\mathrm{cv}(\Delta\bar{E})^2}\right].$$

Where the sampling distribution of the ICER is skewed, this confidence interval will not be symmetrically positioned around the point estimate. This method has been criticized on the grounds that the assumption of joint normality may be hard to justify where sample sizes are small.[5]

2.2. Bootstrap approaches to estimating the ICER confidence interval

Given the unknown nature of the ICER's sampling distribution, there is reason to be cautious of the parametric approaches to confidence interval estimation. A number of commentators have suggested the non-parametric approach of bootstrapping as a possible method of estimating confidence limits for the ICER.[1,2,6,7] The advantage of such intervals is that they do not depend on parametric assumptions concerning the sampling distribution of the ICER.

The bootstrap approach for the simple one sample case is straightforward. Suppose a particular population has a real but unobserved probability distribution F from which a random sample $\mathbf{x}$ of n observations is taken, and the statistic of interest $s(\mathbf{x})$ is calculated. The concern of inferential statistics is to make statements about the population parameter θ based on the sample drawn from that population. In the 'bootstrap world', the observed random sample $\mathbf{x}$ is treated as the empirical estimate of F by weighting each observation in $\mathbf{x}$ by the probability $1/n$. Successive random samples of size n are then drawn from $\mathbf{x}$ *with replacement*† to give the bootstrap samples (re-samples from the original sample). The statistic of interest is calculated for each of these samples and these bootstrap replicates of the original statistic make up the empirical estimate of the sampling distribution for that statistic. This estimated sampling distribution can be used in a variety of ways to construct confidence intervals.

In principle, the bootstrap estimate of the ICER sampling distribution can be obtained in a very similar way to that of the simple one sample case. However, since the ICER is estimated on the basis of four estimators from two samples (equation (1)) care must be taken to bootstrap each sample appropriately. For data structures which are more complicated than a one sample structure, Efron and Tibshirani advocate that the bootstrap mechanism for the observed data

* The solution formula for a quadratic equation of the form $ax^2 + bx + c = 0$ is $-b \pm (\sqrt{(b^2 - 4ac)})/2a$.

† Clearly, sampling from $\mathbf{x}$ without replacement would simply yield $\mathbf{x}$ itself. Hence it is the sampling with replacement which provides the variability through the chance that some observations will appear in the bootstrap sample more than once while others will be omitted altogether.

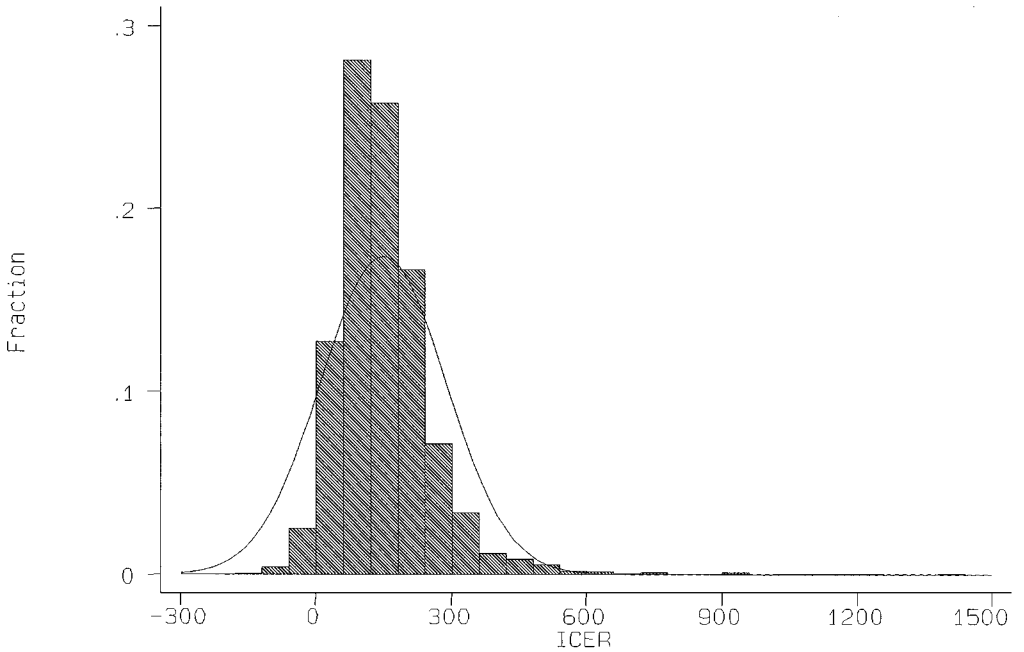


Figure 2. Bootstrap estimation of the sampling distribution of the ICER calculated from clinical trial data (overlaid is a normal distribution with the same mean and variance)

mirror the mechanism by which those original data were obtained.[13] In the case of the ICER, where data on resource use and outcome exists for two groups of patients of size n_A and n_B receiving treatments A and B, respectively, this will involve a three-stage process:

1. Sample with replacement n_A cost/effect pairs from the sample of patients who received treatment A and calculate the bootstrap estimates $\bar{C}_A^*$ and $\bar{E}_A^*$ for the bootstrap sample.
2. Sample with replacement n_B cost/effect pairs from the sample of patients receiving treatment B and calculate the bootstrap estimates $\bar{C}_B^*$ and $\bar{E}_B^*$ for the bootstrap sample.
3. Calculate the bootstrap replicate of the ICER given by the equation

$$R^* = \frac{\bar{C}_A^* - \bar{C}_B^*}{\bar{E}_A^* - \bar{E}_B^*} = \frac{\Delta\bar{C}^*}{\Delta\bar{E}^*}. \tag{7}$$

Repeating this three-stage process many times gives a vector of bootstrap estimates, which is an empirical estimate of the sampling distribution of the ICER statistic. For example, the histogram in Figure 2 shows the estimated sampling distribution from a previously reported study which used the bootstrap to estimate the sampling distribution of the ICER calculated from data generated by an economic evaluation conducted alongside a clinical trial.[8]

Once the sampling distribution of the ICER has been estimated in this way, several approaches exist to estimate confidence limits using the bootstrap estimate of the sampling distribution.

2.2.1. Normal approximation

One method for confidence interval estimation is to take the bootstrap estimate of standard error, given by

$$\hat{\sigma}^* = \sqrt{\left\{\frac{1}{B-1}\sum_{b=1}^{B}(\bar{R}^* - R^{*b})^2\right\}} \tag{8}$$

(where B is the total number of bootstrap replications) and assume that the sampling distribution of the statistic is normal. The resulting $100(1-\alpha)$ per cent confidence interval is

$$(\hat{R} - z_{\alpha/2}\hat{\sigma}^*, \quad \hat{R} + z_{\alpha/2}\hat{\sigma}^*).$$

While comfortingly familiar, this method may be seriously misleading if the sampling distribution is not normal. It ignores the wealth of information in the bootstrap estimate of the sampling distribution, which, as can be seen clearly from Figure 2, may be far from normal.

2.2.2. Percentile

The percentile method avoids this problem by making direct use of the empirical sampling distribution. The $100(\alpha/2)$ and $100(1-\alpha/2)$ percentile values of the bootstrap sampling distribution estimate are used as the upper and lower confidence limits for the ICER. The attraction of this method is its simplicity and its avoidance of the assumption of normality for the ICER. However, it has received considerable criticism from some commentators; for example, Hall (reference 14, p. 36) describes the percentile method as equivalent to '... looking up the wrong statistical tables backwards'.[14] That is, skewed estimation can cause trouble for the percentile method. In particular, in this context, the percentile method assumes that the bootstrap replicates of the ICER are unbiased, whereas it is known that ratio estimators are biased and that bootstrap replicates will magnify the bias of the sample estimate.[15]

2.2.3. Bias-corrected and accelerated

Efron[16] suggests a modification of the percentile method, which seeks to adjust for the bias and skew of the sampling distribution. This is the bias-corrected and accelerated (BCa) percentile method, which involves algebraic adjustments to the percentiles selected to serve as the confidence interval endpoints. The adjusted percentiles are given by

$$\alpha_1 = \Phi\left(\hat{z} + \frac{\hat{z} + z_{\alpha/2}}{1 - \hat{a}(\hat{z} + z_{\alpha/2})}\right)$$

$$\alpha_2 = \Phi\left(\hat{z} + \frac{\hat{z} + z_{(1-\alpha/2)}}{1 - \hat{a}(\hat{z} + z_{(1-\alpha/2)})}\right) \tag{9}$$

where $\Phi(.)$ is the standard normal cumulative distribution function and z_α is the 100α percentile point of the standard normal distribution. Two adjustments to the percentiles are incorporated into equation (9): $\hat{z}$ adjusts the sampling distribution for the bias of the estimator, while $\hat{a}$ adjusts for the skew of the sampling distribution. Setting $\hat{a} = 0$ yields the adjustment for bias on the percentiles chosen to serve as endpoints, and is equivalent to the bias-corrected method

advocated by Chaudhary and Stearns:[5]

$$\alpha_1 = \Phi(2\hat{z} + z_{\alpha/2})$$
$$\alpha_2 = \Phi(2\hat{z} + z_{(1-\alpha/2)}). \tag{10}$$

This bias correction, $\hat{z}$, is given by $\hat{z} = \Phi^{-1}(Q)$ where Q is the proportion of bootstrap replicates which are less than the sample estimate, $\hat{R}$. Therefore, if the bootstrap sampling distribution has median $\hat{R}$, $Q = 0{\cdot}5$ which gives $\hat{z} = 0$ and (in the absence of a skew adjustment) the percentiles from equation (10) correspond to those from the straightforward percentile method. However, where the sampling distribution is not centred on $\hat{R}$ a correction is made for this bias. Notice that the non-linear relationship between the z-score and its probability results in the percentile end points being shifted at unequal rates. It is also worth noting that the bias correction adjustment of the BCa method, while not employing distributional assumptions concerning the sampling distribution of the ICER itself, does make use of parametric assumptions concerning the distribution of the observed bias. This reliance on parametric assumptions has been cited as a potential weakness of the BCa method.[17]

The acceleration constant adjusts for the skew of the sampling distribution. Efron and Tibshirani suggest using a jack-knife estimate for $\hat{\alpha}$:[13]

$$\hat{a}^{**} = \frac{\sum_{1=1}^{n}(\bar{R}^{**} - \hat{R}_i^{**})^3}{6\left[\sum_{i=1}^{n}(\bar{R}^{**} - \hat{R}_i^{**})^2\right]^{3/2}} \tag{11}$$

where $\hat{R}_i^{**}$ is the jack-knife replicate of the ICER with the ith observation removed, $\bar{R}^{**} = \sum \hat{R}_i^{**}/n$ for $i = 1$ to n and $n = n_C + n_T$. In terms of the adjustments to the percentiles given in equation (9), in the absence of a bias correction adjustment, the skew adjustment is given by

$$\alpha_1 = \Phi\left(\frac{z_{\alpha/2}}{1 - \hat{a}z_{\alpha/2}}\right)$$
$$\alpha_2 = \Phi\left(\frac{z_{(1-\alpha/2)}}{1 - \hat{a}z_{(1-\alpha/2)}}\right). \tag{12}$$

Equation (11) shows that if the sampling distribution is symmetric, $\hat{a} = 0$ and equation (12) shows that no adjustment to the percentile interval endpoints is made.

2.2.4. Parametric bootstrap

Efron and Tibshirani outline a simulation-based method of confidence interval estimation that they refer to as a parametric bootstrap approach.[13] Notice that from equation (1), the difference in costs on the numerator and the difference in effects on the denominator of the ICER are both simply the difference between two normally distributed variables, the two sample means* The difference of two means is also normally distributed. The parametric bootstrap approach involves using this property of the distribution of the numerator and denominator in combination with the observed means, variance and covariance to estimate the parameters of the sampling distribution of the cost and effect differences. Sampling from each of these two distributions, while allowing for

* They are normally distributed if the sample sizes are large enough to invoke the central limit theorem or if both costs and effects are normally distributed.

the estimated covariance between them, gives an estimate of the ICER. Repeating this process many times generates an empirical estimate of the sampling distribution of the ICER. The $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of this estimated distribution are used as estimates for the upper and lower limits of the confidence interval, as with the percentile method.

2.3. The Monte Carlo simulation experiments

A simulation experiment was designed to test the coverage properties of each method for calculating confidence intervals in terms of the percentage number of times the true parameter falls outside the interval. Recall that a precise $100(1-\alpha)$ per cent confidence interval will contain the true population parameter $100(1-\alpha)$ per cent of the time in repeated sampling. Therefore, the expectation is that in 100α per cent of samples, the true population parameter lies outside of the interval. In deciding the levels of power and significance to accept, analysts trade off between type I and type II errors. If $\hat{\alpha}$, the observed proportion of Monte Carlo trials where the true population parameter lies outside of the interval, is greater than α, too many type I errors are committed. If $\hat{\alpha}$ is less than α, too many type II errors are committed. Clearly, if an analyst has specified an acceptable rate of error in advance, the method employed should deliver that chosen rate of error.

The Monte Carlo experiments employed the same population parameter values for the average costs and average effects of two hypothetical treatments A and B as those used in the experiments conducted by Wakker and Klaassen.[4] The population mean cost for the group receiving treatment A was set at 40,000 and for group B was set at 30,000; the population mean effects for groups A and B were set as 60 and 50, respectively. Hence the population value of the ICER can be calculated as

$$R = \frac{C_A - C_B}{E_A - E_B} = \frac{40{,}000 - 30{,}000}{60 - 50} = \frac{10{,}000}{10} = 1000. \tag{13}$$

However, in the experiments conducted by Wakker and Klaassen, the standard deviations specified for the population parameters were unrealistically low.[4] Recall that the coefficient of variation for a random variable x is defined as $\mathrm{cv}(x) = \sqrt{\mathrm{var}(x)}/\bar{x}$. Employing the standard deviations and population values specified by Wakker and Klaassen[4] suggests that the average observed coefficient of variation on the numerator of the ratio (the difference in costs) in their simulation experiments was 0·12, while the average observed coefficient of variation of the denominator (effect difference) was approximately 0·02. Low coefficients of variation such as these are likely to give a sampling distribution for the ICER that is very close to a normal sampling distribution.[5] However, we believe this is unrealistic and that many economic evaluations will have much higher coefficients of variation on both the numerator and denominator of the ratio leading to sampling distributions which are significantly skewed. For example, the coefficients of variation for the original data on which Figure 2 is based were 0·55 for the numerator and 0·27 for the denominator. As Figure 2 shows, the estimated sampling distribution was far from normal. Hence the standard deviations of the individual population parameters employed in the Monte Carlo experiments were set such that they generated a range of specified levels of coefficient of variation in the numerator and denominator (details of these calculations are given in the Appendix).

The problem with a single Monte Carlo experiment is that it will be valid only for the chosen parameters and conditions set in that experiment. Hence we designed a series of experiments which systematically varied the underlying conditions most crucial to the shape of the ICER

Table I. Overall performance of the different methods across the 480 experiments

Confidence interval	Lower alpha	Upper alpha	Overall alpha	Low error	Upper error	Overall error
Taylor	0·0051	0·0615	0·0665	−0·0199	0·0365	0·0165
Fieller	0·0205	0·0139	0·0524	−0·0045	0·0069	0·0024
Confidence box	0·0019	0·0047	0·0066	−0·0231	−0·0203	−0·0434
Norm approx	0·0047	0·0507	0·0554	−0·0203	0·0257	0·0054
Percentile	0·0185	0·0376	0·0561	−0·0066	0·0216	0·0061
BCa	0·0229	0·0364	0·0593	−0·0021	0·0114	0·0093
BC	*0·0252	0·0416	0·0668	*0·0002	0·0166	0·0168
Paraboot	0·0155	0·0342	*0·0497	−0·0095	0·0092	*−0·0003

* Non-significant at t-ratio < 2, employing an estimated standard error of a proportion of $\text{se}(p) = \sqrt{\{p(1-p)/n\}} = 0{\cdot}0007$ (for lower/upper alpha/error) and 0·0010 (for overall alpha/error)

sampling distribution. Five different correlation coefficients for the covariance between the costs and effects in the two groups were set: −0·90; −0·45; 0; 0·45, and 0·90. Coefficients of variation for the numerator and denominator were independently specified as 10, 20, 30 and 40 per cent. Six sample sizes were tested: 10; 30; 50; 60; 80, and 100. Population cost and effect data are rarely normally distributed; in particular, cost data is often significantly skewed. Hence, we set the underlying cost and effect data for groups A and B generated in the Monte Carlo experiments to be log-normally distributed.

Each experiment involved taking a random sample of values from one of the specified populations described above. On the basis of the values obtained in these samples, confidence intervals were calculated by each of the seven methods described in Sections 2.1 and 2.2. In addition, the straightforward bias-corrected (BC) bootstrap interval, as employed by Chaudhary and Stearns,[5] was estimated by simply ignoring the accelerator adjustment described in Section 2.2.3. The estimated intervals were then compared to the true ICER from equation (13). Where the true value lay outside of the calculated interval, this result was recorded. This process was repeated 1000 times for each experiment. Hence the number of times the true ICER lay outside the interval divided by the 1000 simulations was the estimated alpha level for that experiment. The upper alpha level recorded the number of times the true ICER lay above the interval, the lower alpha recorded the number of times the true ICER lay below the interval, and the overall alpha was the addition of the upper and lower alphas. Varying all of the conditions above represents 480 different experiments (5 correlation coefficients $\times 4$ coefficients of variation for the numerator $\times 4$ coefficients of variation for the denominator $\times 6$ sample sizes) for which eight confidence intervals were calculated, giving a total of 3840 data points.

3. RESULTS

The overall results across the 480 experiments are presented in Table I. For each of the eight methods the estimated upper, lower and overall alpha rates are shown. To aid interpretation 'error rates' are also shown. These are simply the value of $(\hat{\alpha} - \alpha)$, the estimated value of alpha less the nominal value of alpha chosen for the experiments. The nominal value of alpha appropriate for the upper and lower results is 0·025 and for the overall results is 0·05. Each of the estimated

alpha/error rates was tested for significance using the binomial approximation for the standard error of a proportion.* All were significantly different from the nominal levels except for the lower estimate of the BC bootstrap method and the overall estimate for the parametric bootstrap method.

Care must be taken when interpreting the results of the overall error values. Systematic overestimation in one tail of the distribution combined with underestimation in the other tail can lead to a small overall error generated by large upper and lower errors of opposite sign. This effect is most noticeable in the parametric bootstrap method where the overall alpha is not significantly different from the nominal alpha level. However, it is clear that this is a result of the errors in the upper and lower alpha values cancelling each other out. Similar, although not so dramatic, effects are also apparent in the results for the Fieller and Taylor series methods and the normal approximation, percentile and BCa bootstrap methods.

On the basis of the results from Table I, Fieller's method appears to be performing most consistently across experiments, since it has the lowest upper error, the second lowest lower error and the lowest overall error. The BCa and parametric bootstrap performing best of the non-parametric methods and also outperforming the Taylor method and the box approach. However, since these results are based on summing across the 480 separate experiments, the results presented in Table I could potentially mask underlying variation in the estimated errors between experiments if overestimates in some experiments cancel out with underestimates in other experiments. These variations may be systematically related to experimental factors, which would have significant practical importance.

In order to analyse the effect of these experimental factors on the overall accuracy of the eight confidence interval methods, a technique known as response surface analysis was used.[18,19] This is a technique based on simple OLS regression employing dummy variables for each of the methods. Since the technique requires a single dependent variable we constructed a performance index based on the upper alpha value, defined as $-|\hat{\alpha}^{U} - \alpha^{U}|$. The negative sign was included for interpretative purposes – the greater the number, the better the performance. The upper alpha value was chosen partly due to the problem with the overall alpha value detailed above, but mainly due to the fact that economic analysts are more interested in deciding whether an observed ICER is below some threshold value used for decision making and might therefore be more interested in the upper alpha value being close to its chosen nominal level.[4]

The results of the response surface analysis are presented in Table II. The reference interval chosen was the Fieller method since it seemed to perform best from the results presented in Table I. The natural log of the sample size was chosen as an explanatory variable since it was hypothesized that performance would improve with sample size asymptotically to some limit. The seven confidence interval dummies are presented on separate rows, with the first row being the reference (Fieller) interval. The majority of the coefficients were significant at the standard levels, indicating that there are important differences between the Fieller interval and other methods across the different experiments.

The key to interpreting the results of the analysis lies in the sign of the coefficient; positive coefficients indicate an improvement in performance relative to Fieller's method for that variable and negative coefficients indicate worsening performance relative to the Fieller method. However,

* The binomial approximation for the standard error of a proportion p is given by $\mathrm{se}(p) = \sqrt{\{p(1-p)/n\}}$ where n is the sample size.

Table II. Results of the response surface analysis: estimated coefficients*

Dummy variables	Intercept	Interactions			
		Log of sample size	Coefficient of variation for numerator	Coefficient of variation for denominator	Correlation
Ref(Fieller)	−0·0217†	0·0035†	−0·0106†	0·0108†	0·0013†
Taylor	−0·0159	0·0027†	0·0433	−0·1362†	0·0159†
Confidence box	0·0043†	−0·0041†	0·0088†	−0·0123†	−0·0089†
Norm approx	−0·0338†	0·0052†	0·0299†	−0·0487†	0·0142†
Percentile	−0·0257†	0·0051†	−0·0134†	0·0178†	0·0031†
BCa	−0·0198†	0·0044†	−0·0101†	0·0079†	0·0012†
BC	−0·0348†	0·0069†	−0·0200†	0·0192†	0·0056†
Para boot	−0·0147†	0·0031†	−0·0024	0·0056	0·0001

* Adjusted R^2 for the model = 0·74, $n = 3840$
† t − ratio > 2

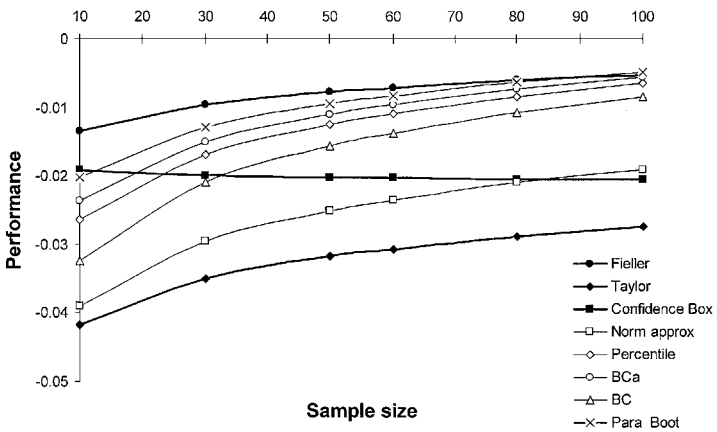


Figure 3. Predicted effect of sample size on the performance variable using results of the *response surface analysis* (coefficients of variation = 0·2, correlation = 0). Performance variable is defined as $-|\hat{\alpha}^U - \alpha^U|$

due to the different intercept values, it is not easy to see the relative performance of each method. In order to demonstrate better the relative performance, the results of the response surface analysis presented in Table II were used to generate predicted performance values for each of the methods. By holding three of the four quantitative variables constant, it was possible to examine the effect of the fourth on the performance of each method.

Figure 3 shows the predicted performance of each method for increasing sample size between 10 and 100, holding the coefficients of variation constant at 0·2 and the correlation coefficient constant at 0. The parametric methods are shown with the weightier lines and solid symbols. At low sample sizes, Fieller's method performs best and the Taylor series performs worst. The confidence box approach appears largely unaffected by sample size. All the other methods

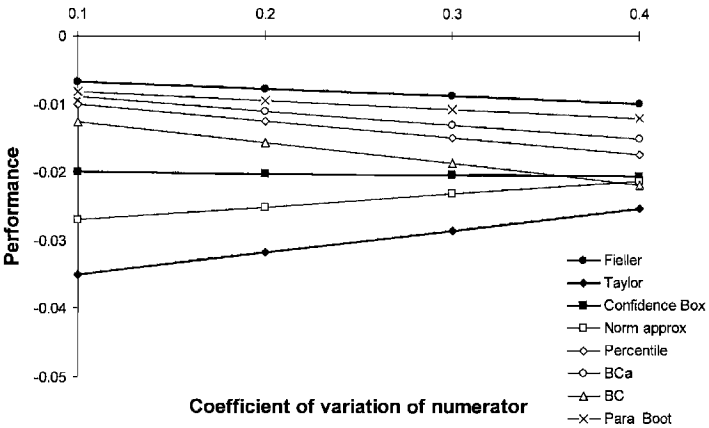


Figure 4. Predicted effect of coefficient of variation of the numerator on the performance variable using results of the *response surface analysis* (sample size = 50, coefficient of variation of the denominator = 0·2, correlation = 0). Performance variable is defined as $-|\hat{\alpha}^{U} - \alpha^{U}|$

improve with sample size, and by $n = 100$ there is little to choose between Fieller's method and the bootstrap methods with the exception of the normal approximation bootstrap method, which, like the Taylor series method, performs poorly.

Figure 4 shows the predicted performance of each method for values of the coefficient of variation of the numerator between 0·1 and 0·4, holding the coefficient of variation of the denominator constant at 0·2, the correlation coefficient constant at 0 and the sample size constant at 50. Again, Fieller's method performs best for all values of the coefficient of variation of the numerator and the Taylor series method performs worst. The normal approximation and Taylor series methods improve in performance as the coefficient of variation of the numerator increases while the performance of the other methods decrease, with the exception of the confidence box method which again appears largely unaffected by changes in the coefficient of variation of the numerator.

Figure 5 shows the predicted performance of each method for values of the coefficient of variation of the denominator between 0·1 and 0·4, holding the coefficient of variation of the numerator constant at 0·2, the correlation coefficient constant at 0 and the sample size constant at 50. A similar picture emerges in that Fieller's method performs best overall, Taylor series performs worst for all but the lowest coefficients of variation of the denominator and the confidence box method seems largely unaffected. This time, however, the performance of the Taylor series and normal approximation bootstrap methods worsen as the coefficient of variation of the denominator increases while the other methods improve in performance.

Figure 6 shows the predicted performance of each method for values of the correlation coefficient between $-0{\cdot}9$ and $+0{\cdot}9$, holding the coefficients of variation constant at 0·2 and the sample size constant at 50. Fieller's method again performs best and appears unaffected by variation in the correlation coefficient of the underlying data. All methods improve with increasing correlation with the exception of the confidence box method, which worsens dramatically as correlation increases. At the very highest correlation, the confidence box method performs worse than the Taylor series method.

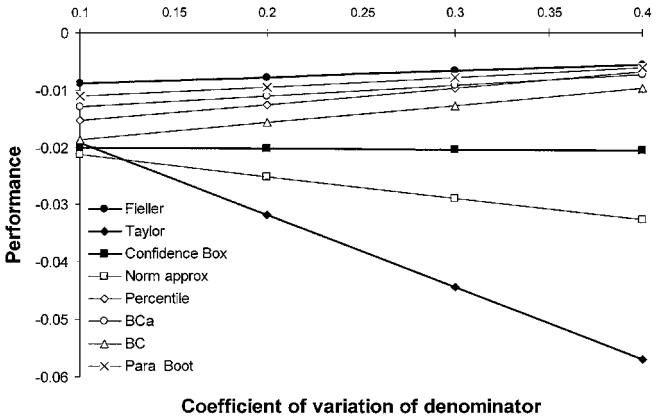


Figure 5. Predicted effect of coefficient of variation of the denominator on the performance variable using results of the *response surface analysis* (sample size = 50, coefficient of variation of the numerator = 0·2, correlation = 0). Performance variable is defined as $-|\hat{\alpha}^{U} - \alpha^{U}|$

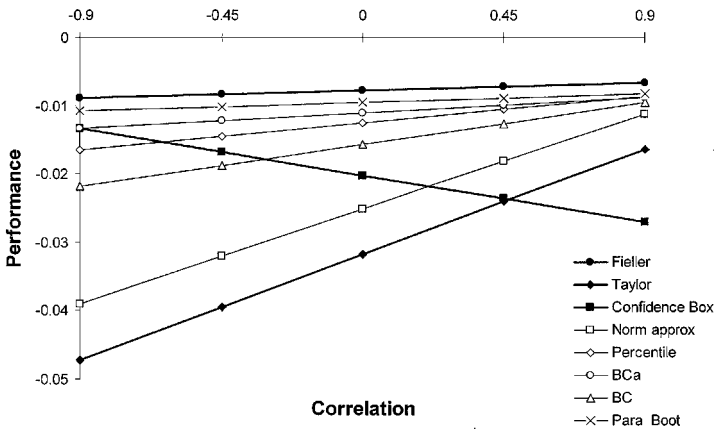


Figure 6. Predicted effect of correlation in the underlying data on the performance variable using results of the *response surface analysis* (sample size = 50, coefficients of variation = 0·2). Performance variable is defined as $-|\hat{\alpha}^{U} - \alpha^{U}|$

4. DISCUSSION AND CONCLUSIONS

The purpose of this paper was to compare a number of parametric approximations of the confidence limits around the incremental cost-effectiveness ratio with non-parametric bootstrapping methods. By devising a series of experiments that represented a realistic range of statistical conditions, we are able to make general conclusions about the relative performance of these approaches, and the factors affecting their relative performance.

No single method dominated (or was dominated by) all other methods across all of the experiments. However, as is shown by Figures 3–6, Fieller's method consistently performed well under a wide variety of assumptions, including small sample sizes, where its assumption of joint

normality between the cost and effect differences has been questioned.[5] Of the bootstrap methods, a clear pattern emerged in terms of the rank ordering of performance. The parametric bootstrap performed the best under most circumstances, closely followed by the BCa and then the straightforward percentile method. The normal approximation method performed most poorly of the bootstrap methods. It is clear that the 'accelerator' adjustment presented by Efron as a refinement to the straightforward bias corrected approach does improve the performance of the method. Although the term 'parametric bootstrap' has been adopted by Efron and Tibshirani,[13] this method is simply a straightforward Monte Carlo simulation of the numerator and denominator of the ratio on the basis of parametric assumptions and the observed means and variances of the data. To what extent this constitutes 'bootstrapping' as the term is commonly applied is an open question.

The predicted effect of increasing sample size is of particular interest. Both the parametric and bootstrap methods rely on asymptotics, and it appears from Figure 3 that the asymptotics of the parametric methods come in to play more quickly than those of the bootstrap methods. In many ways this is a surprising result since bootstrapping has often been linked to the analysis of small samples where standard parametric assumptions are thought to be violated.[17]

As the correlation between the cost and effect in the underlying data increased, the performance of the confidence box method worsened. This is due to the fact that the combination of limits for the confidence box approach is consistent with an assumption of perfect negative covariance between cost and effect. For all other methods, performance increased with increasing correlation. Although there was little to choose between the methods when correlation was high, in practical application it would be unusual to observe extremely high positive or negative correlations. In the data from which Figure 2 was generated, the correlation between cost and effect in the treatment arm of the trial was 0·19, while in the control arm of the trial it was $-0{\cdot}05$.[8] These figures translate into very little covariance between the numerator and denominator of the ICER.

The predicted effect of the coefficient of variation of the numerator and denominator was interesting in that each seemed to influence the methods in the opposite direction. The methods based on an assumption of the normal distribution worsened as the coefficient of variation of the denominator increased, but improved as the coefficient of variation of the numerator increased. For the other methods, the converse was true, with the exception of the confidence box method, which seemed largely unaffected by either coefficient of variation.

One very clear result from these experiments was the inadequacy of methods based principally on the assumption of a normal sampling distribution. Both the parametric based Taylor approximation method and the bootstrap normal approximation were consistently poor performers, although the bootstrap normal approximation seemed to outperform the Taylor series method in general. It is our belief that the sampling distribution of the ICER will almost certainly exhibit an element of skewness in most practical applications, which makes the normal distribution assumption rather limiting.

Recent reviews of economic evaluations have suggested that many authors present only point estimates of cost-effectiveness without any representation of the uncertainty associated with their estimates,[20,21] which suggests that any method of interval estimation is preferable to point estimates alone. However, we have shown that there are substantial differences in the accuracy of the methods advocated in the recent health economics literature. We believe that the nominal error rates accepted by analysts when calculating confidence intervals should be reflected by the actual rates of error that would occur in repeated application of the method. Of course, these error rates will only occur if in practice analysts begin to test hypotheses on the basis of the results

of prospective economic evaluation. We believe the time is now ripe for an analysis of the role of hypothesis testing in economic appraisal.

APPENDIX

This Appendix lays out the method for generating variances of the cost parameters; the same method also applies to effects. Suppose the underlying population cost parameters for treatments A and B are known to be C_A and C_B, respectively. Define the mean difference in cost between two groups sampled from these populations as $\Delta\bar{C} = \bar{C}_A - \bar{C}_B$. In terms of the Monte Carlo experiments, we want to set the coefficient of variation of the difference in costs, $\text{cv}(\Delta\bar{C})$, since it is this which is assumed (in tandem with the effect difference) to determine the shape of the ICER sampling distribution. Hence, the problem is to work backwards from this coefficient of variation to define values for the population cost variance for patients receiving treatments A or B, σ_A^2 and σ_B^2, which will generate the desired $\text{cv}(\Delta\bar{C})$.

We know that the coefficient of variation for the cost difference is defined as

$$\text{cv}(\Delta\bar{C}) = \sigma_{\Delta\bar{C}}/\Delta\bar{C},$$

hence

$$\sigma_{\Delta\bar{C}} = \Delta\bar{C}\,\text{cv}(\Delta\bar{C}). \tag{14}$$

Assuming that random samples of size n_A and n_B are sampled from the population for treatments A and B, respectively, then the treatment costs in each group should be independent. Thus it is possible to relate the variance of the cost difference to the variances of the underlying treatment and control group cost data:

$$\sigma^2_{\Delta\bar{C}} = \frac{\sigma^2_{CA}}{n_A} + \frac{\sigma^2_{CB}}{n_B}. \tag{15}$$

Combining equations (14) and (15) gives

$$[\text{cv}(\Delta C)\,\Delta\bar{C}]^2 = \frac{\sigma^2_{CA}}{n_A} + \frac{\sigma^2_{CB}}{n_B}$$

and rearranging

$$n_A n_B[\text{cv}(\Delta C)\,\Delta\bar{C}]^2 = n_B\sigma^2_{CA} + n_A\sigma^2_{CB}. \tag{16}$$

Further suppose that the coefficients of variation of the underlying costs are the same, that is, $\sigma_{CA}/C_A = \sigma_{CB}/C_B$ or equivalently that

$$\sigma_{CA} = \sigma_{CB}\frac{C_A}{C_B}. \tag{17}$$

Combining equations (16) and (17) gives

$$\begin{aligned} n_A n_B[\text{cv}(\Delta C)\,\Delta\bar{C}]^2 &= n_B\sigma^2_{CB}\left(\frac{C_A}{C_B}\right)^2 + n_A\sigma^2_{CB} \\ &= \sigma^2_{CB}\left[n_A + n_B\left(\frac{C_A}{C_B}\right)^2\right] \end{aligned}$$

and rearranging

$$\sigma^2_{CB} = n_A n_B \frac{[\mathrm{cv}(\Delta C)\Delta\bar{C}]^2}{n_A + n_B(C_A/C_B)^2}. \tag{18}$$

Clearly, σ^2_{CA} can now be calculated from equation (17).

In our experiments, the sample size in each group was the same, that is, $n_A = n_B = n$, therefore equation (18) simplifies to

$$\sigma^2_{CB} = n \frac{[\mathrm{cv}(\Delta C)\Delta\bar{C}]^2}{1 + (C_A/C_B)^2}.$$

ACKNOWLEDGEMENTS

Support of the U.K. Department of Health, the Nuffield College Goodhart fund (AB) and the Office of Health Economics (AB and DW) is gratefully acknowledged. Thanks must go to Dr. James Carpenter for helpful comments on an earlier draft, to Michael Wadley for help transcribing data from the results of the experiments and to the anonymous reviewers who provided valuable input to the final paper. Of course, the views expressed in this document are those of the authors alone, together with the responsibility for any errors.

REFERENCES

1. Mullahy, J. and Manning, W. 'Statistical issues in cost-effectiveness analysis', *in* Sloan, F. (ed.), *Valuing Health Care*, Cambridge University Press, Cambridge, 1994, pp. 149–84.
2. O'Brien B. J., Drummond, M. F., Labelle, R. J. and Willan, A. 'In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care', *Medical Care*, **32**, 150–163 (1994).
3. Willan, A. R. and O'Brien, B. J. 'Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem', *Health Economics*, **5**, 297–305 (1996).
4. Wakker, P. and Klaassen, M. 'Confidence intervals for cost-effectiveness ratios', *Health Economics*, **4**, 373–381 (1995).
5. Chaudhary, M. A. and Stearns, S. C. 'Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial', *Statistics in Medicine*, **15**, 1447–1458 (1996).
6. Mullahy, J. 'What you don't know can't hurt you? Statistical issues and standards for medical technology evaluation', *Medical Care*, **34** (12 Suppl.), DS124–DS135 (1996).
7. Manning, W. G., Fryback, D. G. and Weinstein, M. C. 'Reflecting uncertainty in cost-effectiveness analysis', *in* Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. (eds.), *Cost-effectiveness in Health and Medicine*, Oxford University Press, New York, 1996.
8. Briggs, A. H., Wonderling, D. E. and Mooney, C. Z. 'Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation', *Health Economics*, **6,** 327–340 (1997).
9. Anderson, J. P., Bush, J. W., Chen, M. and Dolenc, D. 'Policy space areas and properties of benefit-cost/utility analysis', *Journal of the American Medical Association*, **255**, 794–795 (1986).
10. Black, W. C. 'The CE plane: A graphic representation of cost-effectiveness', *Medical Decision Making*, **10,** 212–214 (1990).
11. Fieller, E. C. 'Some problems in interval estimation', *Journal of the Royal Statistical Society*, *Series B*, **16**, 175–183 (1954).
12. Cochran, W. G. *Sampling Techniques*, 3rd edn, Wiley, New York, 1977.
13. Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
14. Hall, P. *The Bootstrap and the Edgeworth Expansion*, Springer-Verlag, New York, 1992.
15. Stinnett, A. 'Adjusting for bias in C/E ratio estimates', *Health Economics*, **5**, 469–472 (1996).
16. Efron, B. 'Better bootstrap confidence intervals', *Journal of the American Statistical Association*, **82**, 171–200 (1987).

17. Mooney, C. Z. and Duval, R. D. *Bootstrapping*: *A Nonparametric Approach to Statistical Inference*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-095, Sage, Newbury Park, CA, 1993.
18. Hendry, D. F. 'Monte Carlo experimentation in econometrics', *in* Griliches, Z. and Intriligator, M. D. (eds), *Handbook of Econometrics*, vol. II, Elsevier, Amsterdam, 1994.
19. Mooney, C. Z. *Monte Carlo Simulation*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-116, Sage, Newbury Park, CA, 1997.
20. Udvarhelyi, S., Colditz, G. A. and Epatein, A. M. 'Cost-effectiveness and cost-benefit analyses in the medical literature: are the methods being used correctly?', *Annals of Internal Medicine*, **116**, 238–244 (1992).
21. Briggs, A. and Sculpher, M. 'Sensitivity analysis in economic evaluation: a review of published studies'. *Health Economics*, **4**, 355–371 (1995).

STATISTICS IN MEDICINE
Statist. Med. **18**, 3263–3282 (1999)

EVALUATING THE COST-EFFECTIVENESS OF VACCINATION PROGRAMMES: A DYNAMIC PERSPECTIVE

W. J. EDMUNDS*†, G. F. MEDLEY AND D. J. NOKES

Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL, U.K.

SUMMARY

Although there are many models which are used to calculate the health benefits (and thus the cost-effectiveness) of vaccination programmes, they can be divided into two groups: those which assume a constant force of infection, that is a constant per-susceptible rate of infection; and those which assume that the force of infection (at time t) is a function of the number of infectious individuals in the population at that time (dynamic models). In constant force of infection models the per-susceptible rate of infection is not altered, whereas in dynamic models mass immunization results in fewer infectious individuals in the community and thus a lower force of infection acting on those who were not immunized. We take an example of each of these types of model, examine their underlying assumptions and compare their predictions of the cost-effectiveness of a mass immunization programme against a hypothetical close contact infection, such as measles. We show that if cases of infection are the outcome of interest then the constant force of infection model will always underestimate the cost-effectiveness of the immunization programme except at the extremes when no one or everyone is immunized. However, unlike the constant force of infection model, the dynamic model predicts an increase in the average age at infection after immunization which could impact on the estimate of the cost-effectiveness of the programme if the risk of developing serious disease is a function of the age at infection (as, for instance, is the case for congenital rubella syndrome). Taking cases of infection as the outcome measure and using the dynamic model, the undiscounted cost-effectiveness ratio will tend to decline over time and approach a constant value, as the system moves from pre- to post-immunization equilibrium. We go on to show how the cost-effectiveness of a fixed-term immunization programme might change over time, and discuss why, under most circumstances, decision makers should not assume that elimination (permitting termination of mass immunization) will occur.

1. INTRODUCTION

The resources devoted to health care are limited. Since the market in health care is imperfect[1] some resource allocation decisions must be made. Cost-effectiveness analysis in health care involves the identification of all the relevant alternative uses of a resource (cost) and the evaluation of the expected health gains derived by putting that resource to that use. The aim is to

* Correspondence to: W. J. Edmunds, Immunization Division, PHLS CDSC, Colindale, London NW9 5EQ, U.K.
† Current address: Immunization Division, PHLS CDSC, Colindale, London NW9 5EQ, U.K.

Contract/grant sponsor: Wellcome
Contract/grant number: 040952

CCC 0277–6715/99/233263–20$17.50

maximize the health benefits per dollar spent (or minimize the cost per unit of health benefit gained). Thus cost-effectiveness analysis can be used as an aid to rational public health decision making.[2,3]

There is a large body of literature on both the theoretical and empirical aspects of the transmission dynamics of infectious disease[4–6] which is aimed at both understanding observed epidemiological patterns and predicting the consequences of the introduction of public health interventions to control infection and disease.[7] The approach adopted by these researchers is derived from population biology, and acknowledges the fundamental aspect of infectious disease: that the risk of infection to an individual is related to the number of infectious individuals in the population. In other words, that infectious disease must be considered from a population (or community) perspective, and that individuals cannot be considered in isolation.[8]

Although an important aspect of this work has been directed at design of effective control programmes, it has advanced parallel, and largely separately from, the development of economic analyses of health care interventions.[2,3] The work of Drummond and others has concentrated on improving and standardizing the methods employed to evaluate the cost-effectiveness of health care programmes and has stressed the need for sound epidemiological data on which decisions should be based. One of the consequences of this approach has been the reliance on controlled clinical trials (preferably double blind and randomized) as a source of data for the relative effectiveness of the health care programmes under consideration. Whilst most would not question the use of trial data for estimating the effectiveness of programmes aimed at non-infectious diseases, it has been argued that in general the results of vaccine trials do not provide a good estimate of the effectiveness of a mass immunization campaign.[9–12] Quite simply, most vaccine trials are small relative to a mass programme. Hence, in a well mixed population vaccine trials provide an estimate of the efficacy of the vaccine, that is, the proportion of vaccinated *individuals* who are protected, but not the effectiveness of a mass campaign, that is, the proportion of the *population* who will be protected after mass immunization, and how this might change over time. This is because of the indirect effects of vaccination, that is, the reduced rate of infection in those unvaccinated due to the immunization of a fraction of the community.[4,6,13]

In this paper we argue that these indirect effects (sometimes called herd immunity effects) are important and should be included in a cost-effectiveness analysis of a mass immunization campaign. Hence a different approach is needed from the standard models employed to look at the cost-effectiveness of non-infectious diseases. Note that the discussion centres on the assessment of effectiveness, and we assume that the costs are known. In short, we show that measurement of effectiveness is fundamentally flawed if it does not take a population perspective and include the consequences of interventions directed at one group on the risk of infection on another group (who are not directly targeted, or reached). This is not only because the quantitative details are different, but also because on theoretical grounds effectiveness is being incorrectly assessed. The inclusion of the dynamics of transmission also raises questions that using standard cost-effectiveness methods would not exist. We take the specific example of the consequences of stopping (rather than starting) a vaccination programme to show that transmission dynamics introduces new dimensions into cost-effectiveness studies.

The main body of the paper is taken up by an exposition of the two main techniques for estimating the health benefits derived from mass immunization campaigns: one which derives from the decision analysis literature, and cannot take account of the indirect effects of mass immunization, and one which derives from the ecological literature and does take account of these effects. Prior to introducing the models we define a simple hypothetical example of

a vaccination programme against a close-contact infection[7] which provides the parameter values for both models. The models are introduced and we try to show that the inclusion of population dynamical effects is not prohibitively complicated and does not necessarily make excessive additional data requirements. We then compare the results of the models to illustrate the significance of omission of the indirect effects of vaccination, and go on to introduce and discuss some additional implications of taking a dynamic perspective when evaluating the cost-effectiveness of mass immunization programmes.

1.1. An hypothetical example

In order to compare the two models we first define a basic example which provides the baseline parameter values for both models. We choose a close-contact infectious disease, such as measles or rubella, and, as is customary in such expositions, assume that prior to vaccination the infection has reached endemic equilibrium with respect to time, that is the incidence of infection is not changing over time (ignoring short-term epidemicity[7]). Individuals are assumed infectious for a period of just over a week, after which they become permanently immune and the outcome of interest (that is, the units in which health benefits are measured) is cases of infection. We further assume that there is no appreciable additional mortality associated with infection and that there is a stable population of one million individuals with a mean life-expectancy from birth of 50 years, thus each birth cohort is $1{,}000{,}000/50 = 20{,}000$ in size. Let us assume that a serological survey has revealed that prior to mass immunization 98 per cent of 49 year olds have evidence of past infection. We further assume that vaccination is given at birth and leads to lifelong protection from infection in those individuals who respond to vaccination. Our base-case assumption is that vaccine coverage is complete (that is, 100 per cent) but that only 90 per cent of individuals will be protected following vaccination, although these assumptions are changed in many of the analyses. We assume that the net cost of vaccinating one birth cohort is \$200,000 and that the inflation rate is zero, so the cost of vaccinating subsequent cohorts is also \$200,000 each.

These assumptions are made in part for reasons of simplicity, and in part since they are typical of those encountered in standard decision analysis models (see for instance references 14 and 15). As such we endeavour to keep the example simple and presented in a way suited to the comparison of the two models, and immediately comprehensible to both economists and epidemiologists.

1.2. The cost-effectiveness ratio

The cost-effectiveness ratio is the ratio of the net costs to the net benefits. Immunization programmes are typically implemented over a long period of time, hence the costs and the benefits have to be summed over time, and discounted to their present value (as costs and benefits which occur in the future are valued lower than costs and benefits that occur now). Since we are assuming that the incidence of infection is the outcome of interest, and our decision is to vaccinate or not, then the cost-effectiveness ratio, CER, can be written

$$\mathrm{CER} = \frac{\int C(t)\, \mathrm{e}^{-rt}\, \mathrm{d}t}{\int [\phi^* - \phi(t)]\, \mathrm{e}^{-rt}\, \mathrm{d}t} \tag{1}$$

where $C(t)$ is the net cost of the immunization campaign, $\phi(t)$ is the incidence of infection at time t, ϕ^* is the pre-vaccination incidence (assumed to be constant) and r is the discount rate. There is

some discussion over whether health benefits should be discounted or not; we avoid this issue here and take r to be equal to zero for both health benefits and costs.

2. THE MODELS

Although there are a wide variety of models used in the cost-effectiveness literature, they can be divided into two categories: those that take account of the indirect effects of mass vaccination and those that cannot. The key difference between these classes of models is in their assumptions about the force of infection. The force of infection, λ, is the per-susceptible rate of infection. Thus the incidence of infection equals the force of infection multiplied by the number of susceptibles in the population under study, that is

$$\text{Incidence} = \lambda \times \text{Number Susceptible.} \tag{2}$$

Models which cannot take account of the indirect effects of vaccination, termed here 'constant force of infection' models, treat λ as a fixed parameter. Immunization is assumed to remove a fraction of the susceptible population (that is, reduce the susceptible pool), but not to alter the rate of transmission acting on the remaining susceptibles. Hence cases of infection simply decline in direct proportion to vaccination coverage of the population. Models which can take account of the indirect effects of vaccination, termed here 'dynamic' models, assume λ to be dependent on the number of infectious individuals in the community. Immunization acts by reducing the number of infectious individuals in the population (that is, moving susceptibles directly to the immune class by-passing the infected stage), thereby reducing the force of infection. However (in contrast to constant force of infection models), the proportion susceptible in the population, upon which this new lower force of infection acts, does not necessarily decline from its pre-vaccination level,[4,7] hence incidence does not decrease in direct proportion to coverage.

In the following sections we compare two models, each the simplest from these two categories. Both of them are widely used, well evaluated, and sufficient to show the main differences between the two approaches. The two frameworks model different things, and as such they are not immediately comparable. In the final two sections we therefore incorporate modifications to the dynamic model to enable comparison of like with like.

2.1. The cohort model as an example of a constant force of infection model

The cohort model simulates the effects of the decision (to vaccinate or not) on one birth cohort over the period of their life. The predicted number of infections (the outcome variable in this example) in the cohort is compared with the predicted number of cases in the absence of vaccination. The difference between the two gives the estimate of the health benefits of the programme which is then combined with the cost to give the cost-effectiveness ratio as in equation (1). Note that the reduction in incidence of cases is in simple direct proportion to effective vaccine coverage (that is, level of immunization).

Other than the assumptions of endemic equilibrium prior to immunization and a constant force of infection, an additional assumption must be made by the cohort model, namely the underlying demography of the host population must be modelled in some way so that the number of cases in the cohort over their lifespan can be calculated. Commonly, as in Figure 1, it is assumed that all individuals will live up to their life expectancy (50 years in this case) and then die (this is often referred to as type I survivorship in the ecological literature[4]). Utilizing this extreme

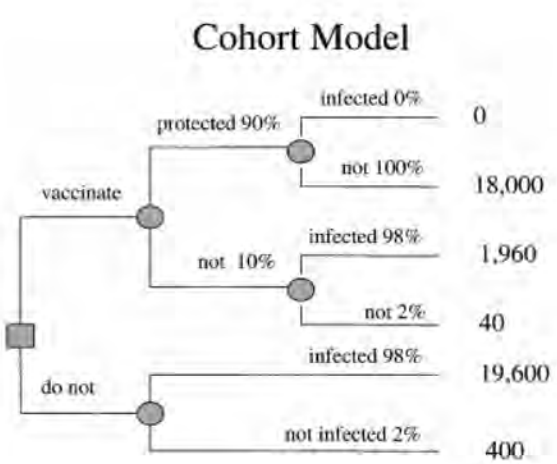


Figure 1. The decision tree is read from left to right. A square node represents a choice node, which denotes a point where the decision maker can elect for one of several courses of action (in this case there is only one decision; to vaccinate or not). A round node represents a chance node, that is, a point at which one of several possible events beyond the control of the decision maker may take place. At each chance node the probabilities must sum to 1. At each chance node the event probability is multiplied by the cohort size, so by the end of the process there are a certain number of individuals at each branch ending. Each birth cohort is 20,000 in size, vaccine efficacy is 90 per cent and type I mortality is assumed

demographic assumption allows the model to be formulated without actually estimating the force of infection explicitly. If any other mortality function is assumed then the force of infection must be explicitly measured, because over time, and age, susceptibles will be lost due to both infection and mortality. In general, ignoring the effect of disease induced mortality, the number of susceptibles in the cohort of age a, $S(a)$, is given by

$$S(a) = S(0) \exp - \int_0^a \lambda(a') + \mu(a')\, \mathrm{d}a' \tag{3}$$

where $S(0)$ equals 20,000 in our example (the initial size of the cohort of susceptibles, that is, at age $a = 0$), $\lambda(a')$ is the force of infection and $\mu(a')$ is the mortality rate. Under the assumption of an age-independent force of infection, and a constant mortality rate (which results in an exponential decline in the number surviving and is often referred to as type II mortality[4]) the above expression reduces to

$$S(a) = S(0) \exp[-(\lambda + \mu)\, a] \tag{4}$$

where λ is the pre-vaccination force of infection and μ is the constant mortality rate. Since the incidence of infection is given by the product of the force of infection and the number of susceptibles (equation (1)) then the cumulative incidence is given by

$$\text{Cumulative incidence(cohort)} = \int_0^L \lambda(a)\, S(a)\, \mathrm{d}a \tag{5}$$

where L is the maximum life expectancy. Substituting the above expression (equation (4)) for the number of susceptibles into this equation and solving the resulting indefinite integral (as for

type II mortality L equals infinity) gives the following expression for the incidence in the cohort using the pre-vaccination equilibrium force of infection, λ^*, with type II mortality:

$$\text{Cumulative incidence(cohort)} = \frac{\lambda^* S(0)}{\lambda^* + \mu} \tag{6}$$

where $S(0)$ is the number of susceptibles at age zero (in our example $S(0)$ equals 20,000 multiplied by the proportion immunized, v).

This expression can then be used to compare the cumulative lifetime incidence in the cohort in the presence of immunization with that in the absence of immunization, the difference between the two being the estimate of the health benefit of the control programme (equation (2)) under the assumption of type II mortality. Type I and type II mortality are simplistic models of population survival. We adopt type II (unless otherwise stated) since it is mathematically convenient and as such is often used in the simplest dynamic model (detailed below).

2.2. A dynamic (SIR) model

This section reviews a simple population dynamical model of the transmission of a microparasitic infection, such as measles, which takes account of the indirect effects of mass vaccination.[4,5,7] The population is divided into three mutually exclusive groups based on their exposure to the infectious agent. These are susceptibles (S), who have not been infected, infectious individuals who have been infected and are currently infectious (I), and immune individuals who have recovered from infection and developed permanent resistance (R) to further infection. In the absence of vaccination, individuals are assumed to be born susceptible. In the presence of vaccination a proportion of individuals, v, are born directly into the resistant class; the remainder are born into the susceptible class. Susceptibles become infected at a time-dependent rate $\lambda(t)$, the force of infection, and pass into the infectious class. Infectious individuals recover and pass into the immune class at a rate γ. Immune individuals remain in this class for the remainder of their lives. Death removes individuals from each of the epidemiological classes at an equal rate, μ (that is, there is no additional mortality due to infection), and total births per unit time equals deaths (from all classes) per unit time, so the population is assumed to remain constant in size. The flow of individuals between the different epidemiological classes (summarized in Figure 2) can be described by the following set of coupled ordinary differential equations (with descriptions alongside):

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N(1-v) - \lambda(t)\,S(t) - \mu\,S(t); \quad \frac{\mathrm{d}S}{\mathrm{d}t} = \text{birth (unimmunized)} - \text{infection} - \text{death} \tag{7}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \lambda(t)\,S(t) - \gamma I(t) - \mu\,I(t); \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \text{infection} - \text{recovery} - \text{death} \tag{8}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu N v + \gamma\,I(t) - \mu\,R(t); \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \text{birth(immunized)} + \text{recovery} - \text{death} \tag{9}$$

where N is the total number of individuals ($N = S(t) + I(t) + R(t)$), and $\lambda(t)$ the force of infection, where

$$\lambda(t) = \beta I(t) \tag{10}$$

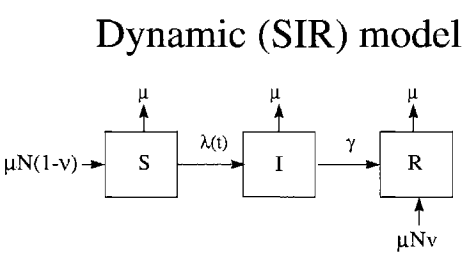


Figure 2. Flow diagram for the basic SIR model. S represents susceptibles, I represents infectious individuals, R represents recovered, or immune individuals and N represents the total population ($N = S + I + R$). The per-susceptible rate of infection (force of infection) is represented by $\lambda(t)$, the recovery rate by γ, and the birth/death rate by μ. All susceptibles are assumed to be vaccinated at birth. The vaccine efficacy is represented by v. See text for details of the model

in which β, the transmission coefficient, is a composite parameter describing the rate at which infectious individuals meet susceptibles and the probability of infection following such a meeting.[4]

The initial conditions for equations (7)–(9) are taken as the pre-vaccination endemic equilibrium number of susceptibles, infecteds and recovereds, which are obtained by setting equations (7)–(9) to zero and solving the resulting set of simultaneous equations.[4] Note that in all simulations presented here, the system was initially set to the endemic equilibrium, thus in the absence of any perturbations, such as immunization, the incidence of infection would remain constant through time.

As stated earlier, the outcome of interest is the incidence of infection. The incidence at time t is given by equation (1), and from equation (10) can be written as $\beta I(t)S(t)$. The cumulative incidence of infection in the entire population from time $t' = 0$ to $t' = t$ is simply given by

$$\text{Cumulative incidence} = \int_0^t \beta I(t')\, S(t')\, \mathrm{d}t'. \tag{11}$$

Thus we can calculate the difference in the cumulative net incidence of infection in the presence and the absence of immunization.

2.2.1. Behaviour of simple SIR model

The dynamic behaviour and the equilibrium (steady state) properties of the SIR model have been explored in detail elsewhere.[4] This section is limited to a discussion of the basic reproduction number, R_0, and an introduction to the notion of stable equilibria. The dynamic behaviour of the model was explored using the software package ModelMaker for Windows Version 3.0.1 (Cherwell Scientific). Note however that, as far as we are aware, there is no commercially available package capable of solving more complicated age- and time-dependent models. These models need to be programmed using an appropriate language, such as FORTRAN, Pascal, or C++.

The basic reproduction number for a microparasitic infection is defined as the average number of secondary infections produced when an infectious individual is introduced into a fully susceptible population. In mathematical terms (based on the model structure of equations

(7)–(10) R_0 can be written as

$$R_0 = \frac{\beta N}{(\mu + \gamma)}. \tag{12}$$

That is, R_0 is the average number of effective contacts made by one infectious individual ($I = 1$) in a population entirely of susceptibles ($S = N$) per unit time, that is βN, over a period of time equal to the average infectious period ($1/(\mu + \gamma)$) of that case.

A related quantity is the effective reproduction number $R_e(t)$, which is defined as the average number of secondary cases generated by a single case at time t, and is given by the product of R_0 and the *proportion* susceptible at time t, $s(t)$. Clearly if $R_e(t) > 1$ then incidence will be increasing and if $R_e(t) < 1$ then incidence will be decreasing. At the specific point when $R_e(t) = 1$, the proportion of the population is at a threshold, often termed s^*. If the proportion susceptible is greater than this threshold then incidence will increase, tending to reduce the number of susceptibles back to this threshold. Similarly, if $s(t) < s^*$, then incidence will decline, enabling the build-up of susceptibles (for example, by birth) to the threshold. Thus the system is dynamically stable. If the proportion of susceptibles is maintained below the threshold by immunizing greater than $1 - s^*$ of the population, then incidence will continue to decline, eventually resulting in elimination of the pathogen from the population.

2.2.2. Estimation of parameters

All the necessary information for estimating the parameters of the SIR model has been provided (see earlier). The mortality rate, μ, equals 1/50 per year, the population size, N equals 1,000,000, γ equals 50 per year and v equals 0·9. It is worth mentioning here that the only additional information needed for the dynamic model with respect to the cohort model is the average infectious period ($1/\gamma$). Although the information is provided, the force of infection, and therefore its constituent parameter β, are not explicitly measured in the simple cohort model with type I mortality, outlined above.

The pre-vaccination equilibrium force of infection, λ^*, can be estimated from serological data by standard methods (see references 4, 13, 16 and 17 for details). Utilizing the simplest of these methods,[17] and assuming that everyone is susceptible at birth and 98 per cent have evidence of infection at 50 years of age, yields an estimate of λ^* of 0·0782 per year. The transmission coefficient, β, can now be estimated as $\lambda^* = \beta I^*$ (from equation (10) at equilibrium), and I^* can be obtained from the steady state solution of equations (7)–(9). Utilizing the base-case parameter assumptions, β can be estimated to be 0·000246 per year. The parameter estimates are summarized in Table I. Note that the use of hypothetical parameter values does not result in loss of generality.

2.3. Modification of the models to enable comparison between them

In order to compare the cohort and dynamic models like with like, we take two approaches. In the first we modify the standard SIR model to monitor the incidence of infection within a single birth cohort of individuals throughout their lifetime. Vaccination is applied, at a defined level, to this and only this single cohort, with zero vaccination coverage in earlier and later cohorts. At first glance this appears to be exactly what is measured by a cohort model. The aim here is to illustrate as a first step the intrinsic differences between the two types of model. We then take a second step and assume that vaccination of a fraction of all births has been maintained for many

years (as is normal in vaccination programmes) until a new steady state is achieved, that is, a new, lower, equilibrium force of infection has been established. We then estimate the cumulative lifetime incidence of infection in a single birth cohort. It should be understood that this too is directly comparable with the cohort model. For in a cohort model the cumulative lifetime incidence in a single cohort is unaltered whether this cohort alone experiences vaccination, or continuous vaccination is in place, since there is no effect of vaccination upon the force of infection. Here the aim is to demonstrate the full magnitude of the indirect effects of vaccination which cannot be captured using the cohort model.

2.3.1. Vaccinating a single cohort in the dynamic model

In order to follow the incidence of infection in a single vaccinated birth cohort, the system of equations is extended to include another susceptible class, S_2, and another infectious class, I_2, which are identical in nature to the other susceptible and infectious classes, S_1 and I_1 (that is, they mix with themselves and all other individuals at the same rate, recover at the same rate and die at the same rate) except that they happen to be born during a specific time period. When vaccination is switched on, individuals are channelled down the susceptible and infectious route defined by S_2 and I_2. When vaccination is switched off, individuals are born into the normal susceptible class once again, and are channelled down this route (that is, S_1 and I_1). Thus if vaccination is switched on for a year, then we can follow the incidence of infection in that year's birth cohort, as well as calculate the incidence in the population as a whole. The equations describing the modified SIR model are described below:

$$\frac{dS_1}{dt} = \mu N(1-\kappa) - \lambda(t)S_1(t) - \mu S_1(t) \tag{13}$$

$$\frac{dS_2}{dt} = \mu N\kappa(1-v) - \lambda(t)S_2(t) - \mu S_2(t) \tag{14}$$

$$\frac{dI_1}{dt} = \lambda(t)S_1(t) - \mu I_1(t) - \gamma I_1(t) \tag{15}$$

$$\frac{dI_2}{dt} = \lambda(t)S_2(t) - \mu I_2(t) - \gamma I_2(t) \tag{16}$$

$$\frac{dR}{dt} = \mu\kappa N v + \gamma[I_1(t) + I_2(t)] - \mu R(t) \tag{17}$$

where

$$\lambda(t) = \beta[I_1(t) + I_2(t)] \tag{18}$$

and k is a switching function which takes the value of 0 when vaccination is switched off and 1 when vaccination is switched on. Thus before vaccination is implemented (and $k = 0$) equations (13) to (18) collapse to equations (7) to (10). When vaccination is switched on (and $k = 1$) a proportion, v, of individuals are born directly into the immune class (as before) whereas the remainder $(1 - v)$ are born into the second susceptible class, S_2.

The cumulative incidence of infection over their entire lifespan in the vaccinated cohort(s) of individuals is given by substituting $S_2(t)$ into equation (11) and the cumulative incidence in the entire population is given, as before, by equation 11 where $S(t) = S_1(t) + S_2(t)$, for limits $t = 0, L$. Note that as we are assuming type II mortality L, the maximum life expectancy, is infinity.

2.3.2. Continuous vaccination in the dynamic model: steady state properties

The SIR model can be reformulated to generate the number of susceptibles, infected and immunes with respect to age under the assumption of time-dependent endemic equilibrium with continuous vaccination in place. That is we are assuming that a vaccination programme has been under way for some time and a new steady state has been achieved in which all cohorts have experienced direct vaccination. Under these circumstances we can set time-dependencies to zero in equations (7)–(10), and instead differentiate with respect to age from birth to life expectancy, L. The initial conditions for this model are now the number of susceptibles, infecteds and immunes at age 0. In our example, prior to immunization, $S(0)$ equals 20,000 and $I(0) = R(0) = 0$. As immunization is assumed to be given at birth, then the initial conditions under immunization are: $S(0) = (1 - v)$* 20,00; $I(0) = 0$; $R(0) = v$*20,000 in which, as before, v is the proportion immunized at birth.

Derivation of the new, steady state λ', is detailed in Anderson and May[4] for both type I and type II mortality. We can then use equation (3) to evaluate the proportion susceptible at age a, $S(a)$, for each survivorship. Note that for type I, where $\mu = 0$ in equation (3), this provides the exact comparison with the cohort model (see Figure 1). Based on estimates of the new equilibrium λ', we can in addition calculate the equilibrium average age at infection under a specified coverage of mass immunization.[4]

3. RESULTS

3.1. Cohort model: the importance of demography

Using type I mortality, effective immunization at 90 per cent is predicted to prevent 17,640 cases over the lifespan of the cohort (19,600–1960, see Figure 1). Estimates of the costs of the different alternatives (to vaccinate or not) can be added to the decision tree to yield the cost per case prevented. Using a net annual cost of \$200,000 yields a cost per infection avoided of \$11·34 (equation (1)). Using the same parameter values but type II mortality, the cohort model predicts a cumulative incidence of 15,928 in the absence of immunization, because some people die before they are infected, and a cumulative incidence of 1593 if 90 per cent of newborns are effectively immunized. Thus the model now predicts that the immunization programme would prevent only 14,335 cases, yielding a cost per case avoided of \$13·95.

3.2. Comparison of results of SIR model with cohort model: vaccinating a single birth cohort

Figures 3(a) and (b) show the difference in the cumulative incidence of infection in a single birth cohort (where it is assumed that this cohort alone is vaccinated), as calculated by the cohort model and the modified SIR (equations (13)–(18)) for a range of proportions immunized at birth (note that we no longer assume that the vaccine has an efficacy of 90 per cent). Results are shown for various values of R_0. Figure 3(*a*) shows the absolute difference in cumulative incidence, and Figure 3(*b*) shows the ratio of the predicted cumulative incidence. The results of the model types differ because of the indirect effects of vaccination, that is, in the dynamic model immunization

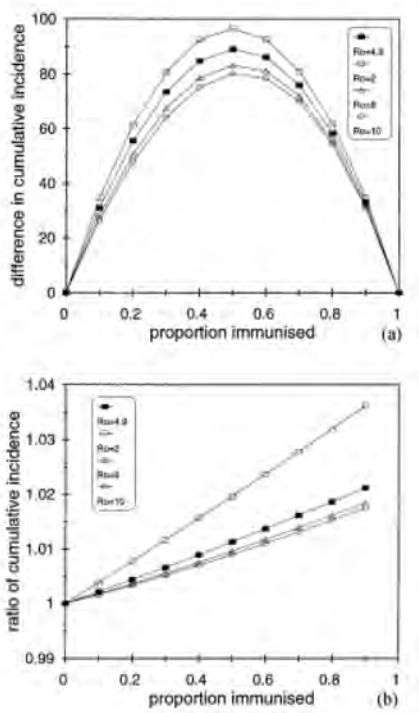


Figure 3. (*a*) The difference in the cumulative incidence in the vaccinated cohort as calculated by the cohort model (equation (6)) and the modified SIR model for different levels of immunization and different values of R_0. Figure 3(*b*) shows the ratio in the cumulative incidence in the cohort calculated by the cohort and modified SIR models. The ratio cannot be calculated if the entire cohort is immunized as both models predict an incidence of zero. In both (*a*) and (*b*) type II mortality is assumed and a single birth cohort is vaccinated. The vaccine efficacy is assumed to be 100 per cent, thus the proportion immunized equals the proportion vaccinated. The baseline parameter set gives an R_0 of 4·9. Other values of R_0 are generated by increasing the transmission coefficient, β, that is all other parameters are held at their baseline parameter values. Altering R_0 means that the initial conditions, that is the equilibrium numbers of susceptibles, infecteds and recovereds, have to be recalculated

leads to a lower force of infection. At the extremes of effective coverage, when the whole cohort is immunized, or no one is immunized, the two models give identical results. When no one is immunized the endemic equilibrium is not perturbed, that is, $\lambda(t) = \lambda^*$ and the two models provide the same results. When the whole cohort is immunized, there is no one left in the cohort to infect, therefore the incidence in the two models is equivalent – both models predict zero incidence. However, at any other level of immunization then the cohort model will always result in an underestimate of the number of cases prevented in the cohort (that is, it will always result in an underestimate of the cost-effectiveness ratio if the outcome of interest is cases of infection). This difference between the two models is greatest (in absolute terms) at intermediate levels of immunization, though in relative terms the difference increases as the proportion immunized increases, as the incidence in the cohort predicted by both models is small at high levels of immunization (Figure 3(*b*)). Note that the cohort model provides a better approximation to the dynamic model at higher values of R_0. This is because at higher values of R_0 the cohort is subject

to higher levels of infection from the other (unvaccinated) cohorts in the population. Thus the cohort model will provide a reasonably close approximation to the dynamic model if the outcome of interest is incidence of infection *in a single birth cohort*, the basic reproductive rate is high, and either very few, or very many within the cohort are immunized.

Note that the differences in the two models are not large (in either absolute, or relative terms). For instance, using the base-case parameter estimates and 90 per cent of the cohort effectively immunized, then the cost-effectiveness ratio as calculated by the cohort model is \$13·95 per case avoided in the cohort and \$13·91 as calculated by the modified SIR model. These small differences are due to the fact that only one birth cohort is immunized – which is clearly unrealistic. Thus although 90 per cent of the cohort are immunized this represents less than 2 per cent of the entire population (each cohort is only 1/50th of the population). In the following section we model the situation in which all cohorts have been vaccinated (and a new equilibrium has been achieved), and in later sections we show how the cost-effectiveness ratio might change over time as the system proceeds to a new equilibrium, that is, when an immunization programme is extended to subsequent birth cohorts.

3.3. Comparison of SIR model with cohort model: steady state under continuous vaccination

Using the estimates of the post-immunization equilibrium force of infection, it is possible to calculate the post-immunization equilibrium average age at infection[4] for different levels of immunization. Figure 4 shows this relationship graphically and compares the results to those in which the force of infection is maintained at the pre-vaccination level (as in the cohort model). It is clear that the constant force of infection model does not predict a shift in the average age at infection, whereas the dynamic model predicts that the average age at infection will increase as immunization coverage increases. It is clear that if the risk of serious consequences following infection changes with age, as, for instance, is the case for congenital rubella syndrome, then this increase in the post-immunization average age at infection may have important repercussions for the effectiveness,[18] and therefore the cost-effectiveness of the immunization campaign.

3.4. The dynamics of cost-effectiveness

In this section we investigate how the effectiveness, and thus the cost-effectiveness, of this simple immunization programme changes over time. Results are for the dynamic model only (equations (7)–(10)) since the cohort model implicity assumes no temporal change. Type II mortality is assumed.

Figure 5 shows a plot of the monthly incidence of infection in the entire population following the introduction of continuous immunization of newborns (beginning year 0) and compares this to the predicted incidence in the absence of immunization. Results will be dependent upon the epidemiological characteristics of the infection (summarized in R_0) and the level of coverage. This example utilizes the baseline set of parameter values (Table I), and effective coverage of 90 per cent. Monthly incidence is predicted to fall and to attain a new lower level over the time period illustrated. The low level incidence is attained after about 2 years of vaccination at this level. The health benefits of the control programme is given by the integral of the difference in the incidence with and without immunization (the denominator of (equation (1)). In the absence of immunization, after 1 year, the predicted number of cases which would be observed is given by the area of the rectangle 0ABC′. In the presence of immunization, the predicted number of cases observed after 1 year is given by area 0ACC′. Therefore after one year of the immunization campaign, the

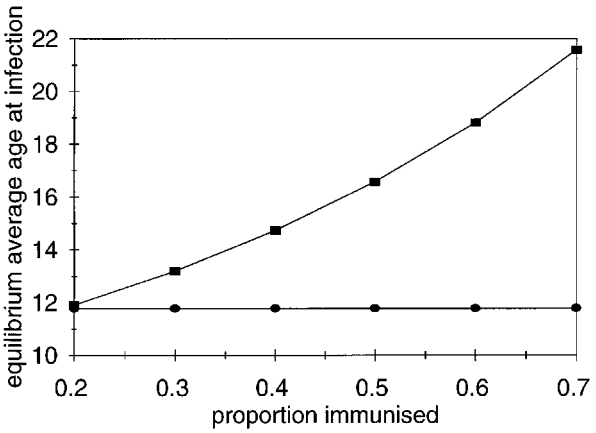


Figure 4. The predicted relationship between the average age at infection at equilibrium and the proportion immunized at birth for, filled circles, a constant force of infection model (cohort model), and, filled squares, the dynamic model (at time dependent equilibrium). Type I mortality is assumed, and vaccination is given at birth

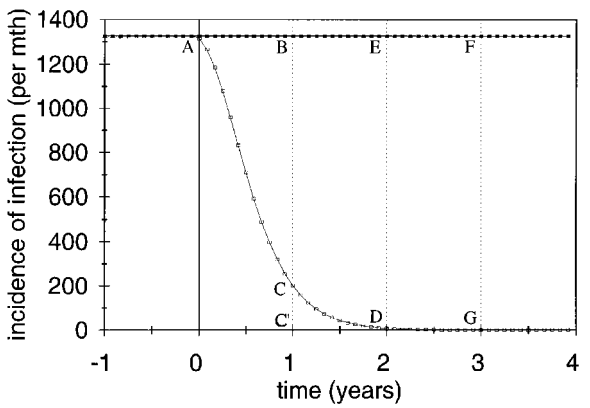


Figure 5. The monthly incidence of infection as calculated by the standard SIR model with (lower line) and without (upper line) mass immunization of infants. The initial conditions are taken to be the endemic equilibrium numbers of susceptibles, infecteds and recovereds. Immunization is implemented at the start of year zero. All infants receive vaccine at birth. The vaccine efficacy is assumed to be 90 per cent. See text for the explanation of the letters

benefit of the programme is given by the area ABC. After two years the total benefit is given by the areas ABC + BCDE, where area BCDE is the incremental benefit of extending the programme for the second year. After three years the total benefit is given by areas ABC + BCDE + DEFG and so on. Notice that area BCDE is larger than area ABC. That is the ratio of the incremental costs to the incremental benefits of extending the programme by one year is smaller than the first year. Thus the programme becomes more cost-effective (more health benefit per $ spent) in the second year. Notice also that area DEFG is similar to area BCDE. Thus the incremental benefit of the third year after immunization is similar to that of the second year (constant marginal

Table I. Parameter estimates used in both models

Description	Symbol	Value
Population size	N	1,000,000
Birth/death rate	μ	1/50 year^{-1}
Recovery rate	γ	50 year^{-1}
Force of infection	λ	0·078 year^{-1}
Transmission coefficient	β	0·000246 year^{-1}
Vaccine efficacy	v	0·9
Net cost		\$200,000 year^{-1}

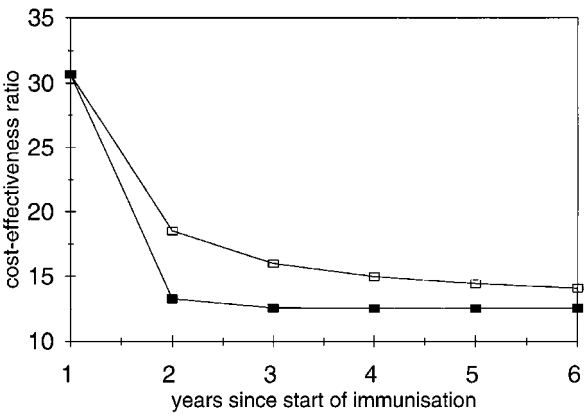


Figure 6. The annual undiscounted incremental (filled squares) and the overall (open squares) cost-effectiveness of the standard immunization programme over time using the standard SIR model with the baseline parameter assumptions (Table I). The effectiveness of the programme is calculated as the cumulative difference in the incidence in the entire population with and without vaccination. The net cost is assumed to equal \$200,000 per year. As in Figure 5, the initial conditions are taken to be the endemic equilibrium numbers of susceptibles, infecteds and recovereds, immunization is implemented at the start of year zero and all infants receive vaccine at birth. The vaccine efficacy is assumed to be 90 per cent

returns). It is clear to see that this pattern will continue as the immunization programme is extended over time as long as stability in incidence is sustained. Figure 6 (filled squares) shows how the undiscounted incremental cost-effectiveness ratio settles at a constant value (of approximately \$12·50 per infection avoided), assuming, as we are, that the net cost of immunization remains constant at \$200,000 per year.

The overall cost-effectiveness ratio of the programme is the total cost divided by the total benefit (equation (1)). Thus as time tends to infinity, the overall cost-effectiveness ratio approaches the incremental cost-effectiveness ratio as the initial non-linear effects of introducing mass vaccination become relatively smaller (open squares, Figure 6). Note that in this example the cost-effectiveness ratio approaches \$12·50 per case avoided, therefore the cohort model results in the programme appearing roughly 12 per cent less cost-effective than the dynamic model.

3.5. Cessation of immunization

The above analysis assumed that once implemented, mass immunization would be continued indefinitely. In this section we use the standard SIR model to assess the implications of termination of an immunization programme.

Figure 7 shows the effect of an immunization programme in which 90 per cent of newborns are immunized between years 0 and 4, after which newborns are left unvaccinated. Figure 7(*a*) is a detail of Figure 7(*b*), that is, it shows the incidence of infection over the first eight years, whereas Figure 7(*b*) shows the same data but extended over a longer period of time. As is evident from Figure 7(*a*), there is a delay following cessation of vaccination before an epidemic ensues. An epidemic occurs when the effective reproductive rate, $R_e(t)$, is greater than 1. That is, the number of susceptibles has to increase (due to births) over a critical value before an epidemic results. The resulting epidemic will eventually settle back at its pre-vaccination equilibrium, when the effective reproductive rate equals one (see Figure 7(*b*)).

4. DISCUSSION

This paper concentrates on methods for estimating the health benefits of vaccination programmes, for which there is no agreed standard method in the economic analysis literature, and consequently there is considerable difficulty in comparison of different cost-effectiveness studies. In particular, the paper compares two methods of estimating the health benefits of a mass immunization programme based on knowledge of the efficacy and likely coverage of the vaccine. There are, however, alternative approaches. One such method is to perform a retrospective study over an adequate time span (determined by the epidemic behaviour of the infection) and estimate the relative risk of infection or disease in the vaccinated and unvaccinated cohorts.[19,20] The data thus obtained will give a direct measure of the health benefits of the programme which includes any indirect benefits, as these will have been experienced by the unvaccinated group. The problems with this technique are that it requires a vaccine to be in use already, the results are not generalizable to a more widespread campaign, and the technique cannot take account of any future temporal changes including any rebound which may occur if the immunization campaign is stopped, that is, it is a static framework. Therefore the evaluation of a new vaccine, or one which is likely to be given to a wider population than that used in the initial study, requires the effects of this vaccine on the population to be modelled in some way.

Unfortunately, as the model must be employed prior to the implementation of the programme under study (if cost-effectiveness analysis is to aid the decision making process), validation of whichever model is chosen will be difficult. Limited validation of the model is provided if it can capture the pre-vaccination trends in incidence and/or prevalence of infection and disease, or if the model is able to capture the post-immunization trends in other epidemiological similar areas, or of a similar infection in the same geographical area. The realistic age-structured SIR model of measles, for instance, has been extensively studied and provides good agreement with pre- and post-vaccination age stratified case reports and seroprevalence surveys.[21]

Whichever model is chosen to estimate the health benefits of an immunization programme, it is clear that the assumptions regarding host demography may critically affect the results, particularly if life-years gained is the outcome of interest, Even in our example, in which cases of infection were the outcome of interest, the cohort model with type I mortality predicted that 90 per cent immunization of infants would lead to 17,640 cases prevented, but only 14,335 cases

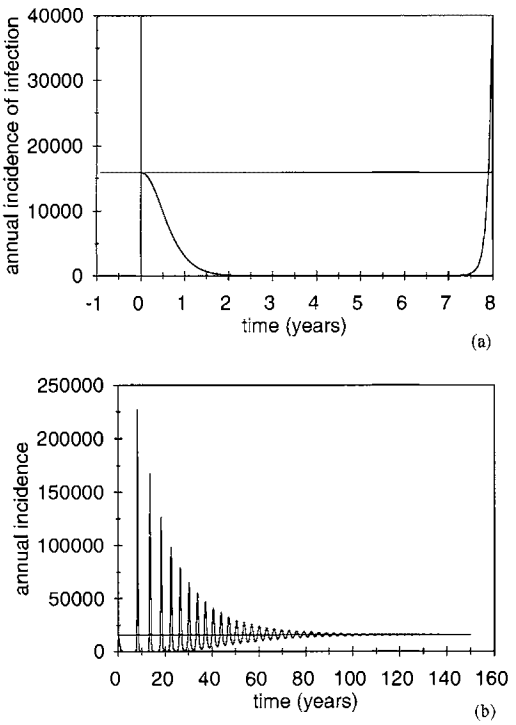


Figure 7. The annual incidence of infection in the entire population as calculated by the standard SIR model. The initial conditions are calculated as before. Mass immunization (all infants born are vaccinated with a vaccine that has an efficacy of 90 per cent) is implemented between years 0 and 4 after which the immunization campaign is halted. Figure 7(*a*) shows a detail, that is, only the first 8 years are plotted, of Figure 7(*b*). The data in the two figures are identical, Figure 7(*a*) merely shows that there is a delay after the mass immunization is halted before an epidemic ensues. Figure 7(*b*) shows the full course of the epidemic

prevented using type II mortality. This is a significant difference in a cohort of only 20,000 individuals. Furthermore, if the infection is rare and localized to groups who, on average, have a life expectancy which is different from the general population, then the demography of these groups should be modelled rather than the demography of the population as a whole.

For instance, in countries with low endemicity of hepatitis B virus, HBV, infection is largely confined to high-risk groups.[22] Excluding mortality due to HBV, individuals in these high-risk groups may experience a lower than average life expectancy due to other infections with which it is associated, for example, the human immunodeficiency virus, or hepatitis C virus. Studies which fail to take account of the additional mortality experienced by these groups may overestimate the effectiveness of vaccination against HBV.

The cohort model is a widely used method for estimating the effectiveness of immunization campaigns in cost-effectiveness analysis. It is based on the decision analysis techniques often used in economic analysis. The technique utilizes data on vaccine efficacy and coverage and multiplies this by the number of individuals in the cohort, to obtain an estimate of the effectiveness of a mass immunization campaign. Thus the model takes an individual perspective and multiplies this up to

give an estimate of the effectiveness of the programme to the population. This approach cannot take account of the indirect effects of mass vaccination. That is, after mass immunization the average force of infection will decrease, and the estimate of the net incidence of infection in the unimmunized fraction of the cohort will in the long term be too high. Analysts often argue that these constant force of infection models lead to an underestimate of the cost-effectiveness ratio, thus if the vaccination programme is deemed to be cost-effective using this framework, then the use of a dynamic model would not change the decision.[23] The obvious problem with this argument arises when the vaccination programme is not deemed to be cost-effective using this technique. Furthermore, it is not always the case that the programme will appear to be more cost-effective using a dynamic model. The decrease in the force of infection following mass immunization can have detrimental as well as beneficial effects. Lowering the force of infection leads to an increase in the average age at infection in the unimmunized portion of the population (Figure 4), as the probability of any one individual coming into contact with an infectious individual is lowered. This increase in the average age at infection has been widely observed following mass immunization.[24,25] If the probability of *disease* following infection is higher in adult age groups (as appears to be the case for many viral infections) then mass immunisation can actually lead to an increase in disease.[18,26,27] Thus it is possible that the health of the population would be decreased by the immunization campaign (negative health benefits), and the programme would not even be deemed 'effective' let alone 'cost-effective'. The constant force of infection model, which the cohort model typifies, cannot take account of these effects.

So when do constant force of infection models provide an adequate description of the proposed immunization programme? There are two situation. First, selective immunization campaigns which are not aimed at 'core groups' (that is a group which contributes significantly to the transmission of the infection) and in which the decision maker is interested in the health benefits which may accrue to the vaccinated group but not others in the population. A good example is provided by Van Doorslaer *et al.*[28] who investigated the cost-effectiveness of vaccination against hepatitis A virus in travellers from low-endemicity areas to high endemicity areas. Second, when the vaccine protects against disease, but does not affect the transmission of the agent, for example, vaccination against tetanus, an environmental pathogen, or rabies, a wildlife zoonosis. In both these cases the vaccine does not reduce the circulation of the pathogen,[6] hence the proportion of individuals protected will equal the proportion of the study population protected. In sum then, constant force of infection models will provide good estimates of the health benefits of a vaccination programme when the force of infection is not substantially altered by mass immunization and/or when the decision maker is interested in a subset of the population (only those liable to be vaccinated).

Apart from allowing the indirect effects of vaccination to be modelled, dynamic models allow an investigation of how the cost-effectiveness of the immunization campaign might change over time. In general, if, prior to immunization, the infection had reached dynamic equilibrium, then the cost-effectiveness ratio will tend towards a constant value. This is because the system will tend to move to a new equilibrium, thus the estimate of the incremental health benefit (of extending the programme through time) becomes the difference between these two equilibrium values and the non-linear effects which occur as the system moves to this equilibrium become relatively less important over time. It is important to stress, however, that there may be a considerable lag between the implementation of mass immunization and a decline in incidence of disease, particularly if the outcome of interests are the sequelae of chronic infection. Thus Edmunds *et al.*[29] have shown that it may take many decades for the incidence of liver cancer due to HBV

infection to approach a constant value. If health benefits are discounted at a positive value then these lags may have a large bearing on the cost-effectiveness ratio.

The importance of taking a dynamic perspective when investigating the health benefits of a mass immunization campaign is highlighted by the cessation of vaccination (Figures 7(*a*) and (*b*)). It is important to note that this is not just a theoretical consideration as the recent debate over cessation of mass BCG immunization in Europe highlights.[30,31] If mass immunization is stopped then the model predicts that an epidemic will ensue when the number of susceptibles has increased over a critical value. There are a number of important issue arising from this epidemic behaviour. This has important implications for the evaluation of fixed term immunization programmes, that is, one which lasts for a fixed period of time, as the consequences of that immunization programme may last for a longer period of time. Thus, in Figure 7(*b*) those areas above the 'do nothing' line represent a cost of the programme (in terms of extra cases caused by cessation of the immunization programme) and those areas below it constitute an extra benefit of the programme. It is worth noting here that if the inter-epidemic period is short, then the area above the line after cessation of mass immunization approximately equals the area below, so that, ignoring the opportunity cost of stimulating an epidemic, the benefits and costs cancel out and the cost-effectiveness of the immunization campaign can be approximated by the estimate of the cost-effectiveness at the time of cessation. Note, however, that this finding is sensitive to model assumptions and cannot be taken as a general result.

In contrast to constant force of infection models, dynamic models predict that immunization above a critical level, which is less than 100 per cent, will eventually lead to elimination in a community in which there is no immigration or emigration (as we have modelled here). Although the SIR model presented in this paper can provide a value for this critical proportion, it cannot predict *when* elimination will occur. The reason is that it is deterministic in nature, which implies that in the model it is possible to have fractions of individuals. Thus, even though only a fraction of an infectious individual might remain in the population, if vaccination is stopped the pool of susceptibles will build up (due to births) and an epidemic will ensue. This problem can be avoided by using a stochastic model in which individuals occur in discrete, integer, units. Thus, if the number of infectious individuals falls below unity, then the model predicts that the infection is eliminated. The problem with using a stochastic version of the SIR model is that, given time, extinction of the infection will always occur naturally, unless a number of infectious individuals are continually introduced into the population. The introduction of infectious individuals also means, however, that the model will never predict a time at which the vaccination programme should be terminated, because once again the termination of mass immunization will result in the build-up of susceptibles and an epidemic will ensue.

This seems to imply that these dynamic models are fundamentally flawed, as both the stochastic and deterministic models will predict that elimination will never occur, and therefore the immunization programme can never be terminated. However, in reality since economic analyses are usually performed from a national, or sub-national, perspective, elimination cannot be included in the model. An individual country or administrative region, through its immunisation programme, might effectively eliminate an infection from within its borders, but would not be able to terminate its immunization programme until all other countries have also eliminated this infection (eradication). Any one individual country cannot ensure that global eradication would ensue, therefore they should not, when evaluating a national level immunization programme, consider the possibility of elimination (as elimination can only be considered temporary, and

vaccination must be maintained until global eradication is achieved – a phenomenon which is usually beyond the sphere of influence of the decision maker).

The inclusion of a proper population perspective complicates cost-effectiveness studies of immunization programmes. The theoretical reasoning for its inclusion is akin to that for the inclusion of discounting in an economic evaluation; to leave it out on the grounds of complexity is to avoid a fundamental process that is known to occur for the sake of being able to complete an incorrect analysis.

ACKNOWLEDGEMENTS

WJE is in receipt of a Wellcome Post-doctoral Health Services Research Training Fellowship (grant number 040952). DJN is a Royal Society University Research Fellow. We thank the members of WUPERT for useful discussions.

REFERENCES

1. Arrow, K. J. 'Uncertainty and the welfare economics of medical care', *American Economic Review*, **53**, 941–973 (1963).
2. Weinstein, M. C. and Stason, W. B. 'Foundations of cost-effectiveness analysis for health and medical practices', *New England Journal of Medicine*, **296**, 716–721 (1977).
3. Drummond, M. F., Stoddart, G. L. and Torrance, G. W. *Methods for the Economic Evaluation of Health Care Programmes*, Oxford Medical Publications, Oxford, 1987.
4. Anderson, R. M. and May, R. M. *Infectious Diseases of Humans. Dynamics and Control*, Oxford University Press, Oxford, 1991.
5. Nokes, D. J. and Anderson, R. M. 'The use of mathematical models in the epidemiological study of infectious diseases and in the design of mass immunization programmes', *Epidemiology and Infection*, **101**, 1–20 (1988).
6. Fine, P. E. M. 'Herd immunity: history, theory, practice', *Epidemiologic Reviews*, **15**, 265–302 (1993).
7. Anderson, R. M. and Nokes, D. J. 'Epidemiological approaches: mathematical models of transmission and control', *in* Holland, W. W., Detels, R. and Knox, E. G. (eds), *Oxford Textbook of Public Health*, Oxford University Press, Oxford, 1996.
8. Medley, G. F. 'Conflicts between the individual and communities in treatment and control', *in* Isham, V. and Medley, G. F. (eds), *Models for Infectious Human Diseases. Their Structure and Relation to Data*, Cambridge University Press, Cambridge, 1996, pp. 331–343.
9. Halloran, M. E., Haber, M., Longini, I. M. and Struchiner, C. J. 'Direct and indirect effects in vaccine efficacy and effectiveness', *American Journal of Epidemiology*, **133**, 323–331 (1991).
10. Haber, M., Longini, I. M. and Halloran, M. E. 'Measures of the effects of vaccination in a randomly mixing population', *International Journal of Epidemiology*, **20**, 300–310 (1991).
11. Longini, I. M., Halloran, M. E., Haber, M. and Chen, R. T. 'Measuring vaccine efficacy from epidemics of acute infectious agents', *Statistics in Medicine*, **12**, 249–263 (1993).
12. Clemens, J., Brenner, R., Rao, M., Tafari, N. and Lowe, C. 'Evaluating new vaccines for developing countries. Efficacy or effectiveness?', *Journal of the American Medical Association*, **275**, 390–397 (1996).
13. Anderson, R. M. and May, R. M. 'Vaccination and herd immunity to infectious diseases', *Nature*, **318**, 323–329 (1985).
14. Bloom, B. S., Hillman, A. L., Fendrick, A. M. and Schwartz, J. S. 'A reappraisal of hepatitis B virus vaccination strategies using cost-effectiveness analysis', *Annals of Internal Medicine*, **118**, 298–306 (1993).
15. Margolis, H. S., Coleman, P. J., Brown, R. E., Mast, E. E., Sheingold, S. H. and Arevalo, J. A. 'Prevention of hepatitis B virus transmission by immunization. An economic analysis of current recommendations', *Journal of the American Medical Association*, **274**, 1201–1208 (1995).
16. Grenfell, B. T. and Anderson, R. M. 'The estimation of age-related rates of infection from case notifications and serological data', *Journal of Hygiene, Cambridge*, **95**, 419–436 (1985).
17. Anderson, R. M. and May, R. M. 'Vaccination against rubella and measles: quantiative investigations of different policies', *Journal of Hygiene, Cambridge*, **90**, 259–325 (1983).

18. Anderson, R. M. and Grenfell, B. T. 'Quantitative investigations of different vaccination policies for the control of congenital rubella syndrom (CRS) in the United Kingdom', *Journal of Hygiene, London*, **96**, 305–333 (1986).
19. Mullooly, J. P., Bennett, M. D., Hornbrook, M. C., Barker, W. H., Williams, W. W., Patriarca, P. A., and Rhodes, P. H. 'Influenza vaccination programs for elderly persons: cost-effectiveness in a health maintenance organization', *Annals of Internal Medicine*, **121**, 947–952 (1994).
20. Nichol, K. L., Margolis, K. L., Wuorenma, R. N., and Von Sternberg, T. 'The efficacy and cost effectiveness of vaccination against influenza among elderly persons living in the community', *New England Journal of Medicine*, **331**, 778–784 (1994).
21. Babad, H. R., Nokes, D. J., Gay, N. J., Miller, E., Morgan-Capner, P. and Anderson, R. M. 'Predicting the impact of measles vaccination in England and Wales: model validation and analysis of policy options', *Epidemiology and Infection*, **114**, 319–344 (1995).
22. Hollinger, F. B. and the North American Regional Study Group 'Controlling hepatitis B virus transmission in North America', *Vaccine*, **8(suppl)**, s122–s128 (1990).
23. Margolis, H. S., Coleman, P. J. and Mast, E. E. 'Cost-effectiveness of hepatitis-B virus immunization', *Journal of the American Medical Association*, **275**, 909 (1996).
24. Lau, Y., Chow, C. and Leung, T. 'Changing epidemiology of measles in Hong Kong from 1961 to 1990 - Impact of a measles vaccination program', *Journal of Infectious Diseases*, **165**, 1111–1115 (1992).
25. Ukkonen, P. and Von Bonsdorff, C. H. 'Rubella immunity and morbidity: effects of vaccination in Finland', *Scandanavian Journal of Infectious Diseases*, **20**, 255–259 (1988).
26. Anderson, R. M., Crombie, J. A. and Grenfell, B. T. 'The epidemiology of mumps in the UK: a preliminary study of virus transmission, herd immunity and the potential impact of immunization', *Epidemiology and Infection*, **99**, 65–84 (1987).
27. Edmunds, W. J., Medley, G. F. and Nokes, D. J. 'Vaccination against hepatitis B virus (HBV) in highly endemic areas: waning vaccine induced immunity and the need for booster doses', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **90**, 436–440 (1996).
28. Van Doorslaer, E., Tormans, G. and Van Damme, P. 'Cost-effectiveness analysis of vaccination against hepatitis A in travellers', *Journal of Medical Virology*, **44**, 463–469 (1994).
29. Edmunds, W. J., Medley, G. F. and Nokes, D. J. 'The transmission dynamics and control of hepatitis B virus in The Gambia', *Statistics in Medicine*, **15**, 2215–2233 (1996).
30. Trnka, L., Dankova, D. and Svandova, E. 'Six years' experience with the discontinuation of BCG vaccination. 2. Cost and benefit of mass BCG vaccination', *Tubercle and Lung Disease*, **74**, 288–292 (1993).
31. Watson, J. M. 'BCG - mass or selective vaccination?', *Journal of Hospital Infection*, **30(Suppl)**, 508–513 (1995).

STATISTICS IN MEDICINE
Statist. Med. **18**, 3283–3298 (1999)

A MONITORING SYSTEM FOR DETECTING ABERRATIONS IN PUBLIC HEALTH SURVEILLANCE REPORTS†

G. DAVID WILLIAMSON[1]* AND GINNER WEATHERBY HUDSON[2]

[1]*Epidemiology Program Office, MS K73, Centers for Disease Control and Prevention, 4770 Buford Highway, Atlanta, GA 30341, U.S.A.*

[2]*Drexel University, College of Business and Administration, Matheson Hall, 32nd and Market Streets, Philadelphia, PA 19104, U.S.A.*

SUMMARY

Routine analysis of public health surveillance data to detect departures from historical patterns of disease frequency is required to enable timely public health responses to decrease unnecessary morbidity and mortality. We describe a monitoring system incorporating statistical 'flags' identifying unusually large increases (or decreases) in disease reports compared to the number of cases expected. The two-stage monitoring system consists of univariate Box–Jenkins models and subsequent tracking signals from several statistical process control charts. The analyses are illustrated on 1980–1995 national notifiable disease data reported weekly to the Centers for Disease Control and Prevention (CDC) by state health departments and published in CDC's *Morbidity and Mortality Weekly Report*. Published in 1999 by John Wiley & Sons, Ltd. This article is a U.S. Government work and is in the public domain in the United States.

INTRODUCTION

Public health surveillance provides the foundation for much of effective epidemiologic and public health practice. It is the ongoing collection, analysis, interpretation and dissemination of outcome-specific information for use in planning, implementing and evaluating public health practice.[1,2] Data collected from public health surveillance systems can provide important clues to the aetiology of a disease and assist in identification of important risk factors, onset of epidemics and detection of unusual observations in reports of infectious diseases and other conditions, thus facilitating early public health response to minimize undue morbidity and mortality.[3–5]

There is a long distinguished history in modelling public health data to provide greater insight into aetiology, spread, prediction and control of diseases. One such early effort was by William Farr when, in 1840, he fit a normal curve to deaths from smallpox in hopes of discovering why epidemics appear and disappear.[6] Other epidemiologic and statistical analyses have focused on modelling disease incidence and prevalence, geographical distribution and spread of disease, and on forecasting disease counts and associated health care needs.[7–17] Models have been developed to detect time and/or spatial clusters of disease and, recently, to smooth rates in small area estimation problems (that is, to overcome statistical issues occurring when the unit of analysis is associated with a geographic area which is small relative to the area spanned by a set of

* Correspondence to: G. David Williamson, Epidemiology Program Office, MS K73, Centers for Disease Control and Prevention, 4770 Buford Highway, Atlanta, GA 30341, U.S.A. E-mail: dxw2@cdc.gov

Published in 1999 by John Wiley & Sons, Ltd.

contiguous geographic areas analysed).[18–23] Because some epidemiologic and, in particular, public health surveillance data are collected in time sequence at regular intervals in an ongoing manner, these data often exhibit correlation, non-stationarity (in the mean and/or variance) and seasonality, characteristics which statistical time series modelling is especially suited to accommodate.[24]

Of critical importance to public health practitioners is an ability to detect quickly substantial changes in disease and other consequential epidemiologic data series, thus facilitating timely public health response to mitigate morbidity and mortality.[25–27] Because of the special features of public health surveillance data as listed above, and including that the data are not usually generated from a random sample, detection of changes in public health data presents an analytic challenge. However, particularly in the last 15 years, several research efforts have described statistical and epidemiologic methods to detect substantial changes in public health data series, including timely signalling of the onset of epidemics or identification of aberrations (that is, statistically significant departures in the occurrence of a health event from what is expected based on the historical incidence of the event) in the data.[27–36] These methods include a bar graph based on the ratio of current to historical data, extensions to the linear dynamic model and applications of the Kalman filter and probability index function. Additionally, Box–Jenkins time series models have been applied to public health forecasting problems in the past[13,37–41] and statistical process control (SPC) methods have been discussed as evaluation tools for public health surveillance,[42] although we are unaware of any previous literature describing a combination of the two methods for a public health monitoring system.

The objective of this research was to develop a monitoring system to (i) detect aberrations in reported data on disease incidence, and (ii) provide a signal to alert public health practitioners to undertake timely public health action. Here we develop and introduce a two-stage monitoring system comprised of two traditional time series techniques, Box–Jenkins autoregressive, integrated, moving average (ARIMA) models and SPC charts, for the detection and signalling of aberrations in public health data. A substantial difference between other techniques and our monitoring system is that we combine the strengths of Box–Jenkins and SPC methods, taking into account the order of the data in time as well as explicitly incorporating methods which recognize and account for the correlation among observations which occur due to the influence of the same (epidemiologic) factors at adjacent or nearby time periods (referred to as autocorrelation in the time series literature).

In this paper, we describe our innovative modelling approach to detecting aberrations and apply the methods to United States disease report series. We consider issues which arise when employing the monitoring system and discuss further directions to pursue for enhancing our ability to detect aberrations in public health surveillance data.

METHODS

Data

Analyses for this investigation were performed on data from the National Notifiable Diseases Surveillance System (NNDSS) of the Centers for Disease Control and Prevention (CDC). The NNDSS database consists of weekly reports of 52 diseases (as of 1 January 1996) designated by the Council of State and Territorial Epidemiologists as nationally notifiable and approved by the state and territorial health departments for reporting to CDC.[43] A notable strength of the

NNDSS is its timeliness. Reports of cases of diseases are aggregated weekly and sent by each state health department to CDC for dissemination in CDC's *Morbidity and Mortality Weekly Report*. Data analysed here are provisional (that is, do not include updates from states as additional information becomes available) to allow for the most timely monitoring of disease series and any necessary subsequent public health action.

Data analyses were performed in two phases, the first beginning in 1990 and the second in 1995. For the original analysis, 17 diseases were chosen to represent a diversity of disease and demographic characteristics, including disease aetiology, seasonality and incidence, population distribution and geography. These diseases were acquired immunodeficiency syndrome (AIDS), aseptic meningitis, encephalitis, gonorrhoea, hepatitis type A, hepatitis type B, hepatitis non-A and non-B, legionellosis, malaria, measles, meningococcal infections, mumps, pertussis, rubella, syphilis, tuberculosis and typhus fever (tickborne). In addition to investigating application of our methods on the national data for these 17 diseases, we applied our modelling strategy and control charts to state data for gonorrhea, hepatitis type A, measles, meningococcal infections and typhus fever (tickborne). We applied the methods to data from each of three states for hepatitis type A, meningococcal infections and typhus fever (tickborne), and to data from each of four states for the other two diseases. These diseases and states were chosen to represent a wide array of disease and demographic characteristics. This first phase included data from week one 1980 to week 52 1989. The second analysis phase augmented the original data with disease reports from week one 1990 to week 19 1995 and has focused to date only on the national disease series for hepatitis type A. Approximately every six years there is a 53rd week of data included in the disease series. For these years, we average the number of reported cases for weeks 52 and 53, creating a new value for week 52 and ensuring consistency in the number of weeks each year for analysis.

Analysis Strategy

Each of our disease report series is a time-dependent phenomenon which is affected by various factors, some known and some unknown. Because of the complexity and unknown nature of some of the factors affecting disease reports, stochastic rather than deterministic models are used to better represent and forecast numbers of reported cases of disease. Many disease series contain trend, seasonal and cyclical components which result in substantial autocorrelation. We employ the Box–Jenkins modelling strategy on the surveillance data because these techniques are designed to properly handle autocorrelation issues.[24] If the ARIMA Box–Jenkins time series models provide an adequate fit to the data, they will produce one-step-ahead minimum variance forecasts of the process. These forecasts are the expected number of reported cases of disease and are based on the historical provisional data. The associated forecast errors (that is, the differences between the model forecasts and the observations) should then be approximately independently and identically distributed (IID).

Statistical process control charts are among the most prevalent and valid methods for monitoring time series data, but their use usually requires observations to be IID random variables when the process is in statistical control.[44] If assumptions for applying Box–Jenkins methods are met and adequate ARIMA models developed, then the subsequent forecast errors from the models can be monitored by the SPC charts. Thus the objective of an appropriate monitoring system is not to track week-to-week changes in the number of reported cases, but rather to identify any substantial variation in the weekly expected number of reported cases as

determined from historical patterns. The control charts employ statistical limits to generate 'flags' identifying deviations from historical data patterns and from the underlying stochastic process generating the observations (that is, the charts detect aberrations). These aberrations signal a potential change in the pattern of incidence of that health event or in the associated reporting procedures, thus facilitating a timely public health response.

Thus we have developed and present here a two-stage monitoring system for public health surveillance data based on the successful integration of (i) ARIMA time series models providing dynamic forecasting of future expected disease reports and (ii) SPC methods for tracking the forecast errors from the ARIMA models. Similar approaches have been suggested for application in manufacturing and industrial settings.[45,46]

Preliminary analysis

We produced several time plots for each of the national and state disease report series we analysed to better understand the historical data and guide the Box–Jenkins and SPC analyses. These graphics included (i) number of reported cases each week (for example, from week one 1980 to week 52 1989 for the first analysis phase) to visually portray the entire series of data; (ii) number of reported cases each week over the last two years of the period of analysis to more carefully depict recent patterns in historical data; (iii) number of reported cases each week plotted vertically by year (that is, each of the 52 weekly reports plotted in a vertical line above the 'year' label on the horizontal axis) and connected at the annual mean to investigate evidence of non-stationarity in the mean level and variability of weekly reports within each year of the series; and (iv) number of reported cases each week plotted vertically by week number, one to 52 (for example, for 1980–1989 analysis, each of the ten week-one reports, one for each year, are plotted in a vertical line above the 'week one' label on the horizontal axis, with data for the other 51 weeks plotted similarly) and connected at the weekly mean to show the range of values for the same week number across years, thus revealing any seasonality in the data.

Development of Box–Jenkins ARIMA time series models for forecasting

In order to apply Box–Jenkins methods to time series data, the series must meet certain stationarity assumptions; the series must be stationary in both mean and variance across the modelling period. A series not stationary in the mean can be made so generally by calculating a differenced series (that is, by taking the differences in number of reported cases of disease between two specified reporting periods, generally successive ones). A series not stationary in the variance of its observations can many times be made so by transforming the data (for example, with a square root transformation).[47,48]

For those data series stationary in both the mean and variance, we applied Box–Jenkins techniques to develop forecasting models.[24,47,48] Often mathematical transformations and/or differencing was required to achieve stationarity. The Box–Jenkins approach to time series forecasting requires an adequate stochastic ARIMA model of the following form to describe the time series under study:

$$\Phi_p(B)\,\nabla^d z_t = \Theta_q(B)\,a_t$$

where $\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ is an autoregressive polynomial of order p, ∇ is the backward difference operator, d is the order of the first difference, z_t is the observation at time t, $\Theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$ is a moving average polynomial of order q, B is the

backshift operator, and a_t is a sequence of normally and independently distributed random 'shocks' with mean zero and constant variance.

To develop and select an appropriate ARIMA model for each series, extensive identification, estimation and diagnostic checking (that is, model evaluation) stages were completed. This work was accomplished using SAS/ETS® software.[49] The modelling procedure for each series began by examining the correlative structure of the data in the autocorrelation and partial autocorrelation functions to identify tentative models. Parameters were then estimated for each tentative model using disease report data for a selected period.

Model residuals, the difference in observed and model-estimated values for the historical period used in model estimation, and forecast errors were examined in the diagnostic checking stage, which was comprised of evaluating statistical fit and model forecasting capabilities. Forecast errors were determined by withholding a period of the most recent reports from the historical data used to estimate the ARIMA model parameters. A comparison of the model forecast with the actual reported cases that had been withheld could then be made and the corresponding forecast errors (actual minus forecast values) calculated. The adequacy of the statistical fit was determined by the statistical significance of any autoregressive (AR) or moving average (MA) terms in the model and by analysis of the model residuals, which should be basically free of autocorrelation and patterns. Model performance for the estimation period and the forecast period was quantified on the residuals and forecast errors, respectively, by calculating various statistics, including mean absolute deviation (MAD) and root mean squared error (RMSE).

In both analysis phases of this work, beginning in 1990 and again in 1995, the historical data were divided into two periods, one for model estimation and one for forecasting, model evaluation and testing. For the 1990 analysis phase, we developed a model for the 1980–1987 historical data and forecasted weekly reported cases for all 52 weeks in 1988, the forecast for the 52nd week of 1988 being termed a 52-week-ahead forecast. Then, to check the robustness of the form of the ARIMA model to changes in the data (that is, to inclusion of additional observations), using the same model form as for the 1988 forecasts, we re-estimated the model coefficients with 1980–1988 historical data and forecasted all 52 weekly values for 1989, ultimately producing a 52-week-ahead forecast for the last week of 1989. Within production of the 1989 forecasts, we did not incorporate the more up-to-date 1989 disease reports and there was no redevelopment of model form or re-estimation of model coefficients after each weekly forecast was calculated.

For the 1995 analysis phase, we developed an ARIMA model for 1980–1993 data, then forecasted weekly values for 1994 and the first 19 weeks of 1995 in two ways. First we developed a model for the 1980–1993 data and forecasted weekly values for all 52 weeks of 1994 and the first 19 weeks of 1995, ultimately producing a 71-week-ahead forecast. In this first approach, we did not incorporate the more up-to-date 1994–1995 disease reports nor was there redevelopment of model form or re-estimation of model coefficients after each weekly forecast was calculated. Secondly, we developed one-week-ahead, rolling forecasts for each of the 71 weeks of 1994 and 1995 by retaining the same model form at each forecasting step, but re-estimating the model coefficients after the addition of each week's newly reported cases from 1994 and first 19 weeks of 1995 to incorporate as much data as were available in the estimation step, and thus produce a more valid forecast. This second approach was consistent with recommendations that Box–Jenkins models produce the most valid results when forecasting one-step-ahead values.[24] This latter strategy was also implemented to simulate how the monitoring system might be

utilized by CDC and state health departments and to test the robustness of the model coefficients to the inclusion of additional data each week by allowing comparison of results from the two manners in which 1994–1995 data were forecasted.

Use of SPC methods and control charts for tracking signals

Statistical process control charts are a major category of methods for creating tracking signals. These methods, dating back to Walter Shewhart's work, employ collecting data sequentially in time and plotting those observations or functions of the observations on control charts.[50] In the basic Shewhart-type chart, a plotted point is compared to predetermined control limits, and if the point falls beyond these critical boundaries, it is a signal that the process is statistically out of control, or that a statistical aberration has been identified. Although calculation of control chart statistics is relatively easy, it is sometimes difficult to determine the most effective control charts and appropriate control limits for the specific monitoring problem.[44]

In the proposed system, we considered the following three types of control charts: (i) Shewhart; (ii) moving average; and (iii) exponentially weighted moving average (EWMA). These charts were implemented using SAS/QC^{R} software.[51]

Upper and lower control limits in the Shewhart control chart are typically set at ± 3 standard deviations from the overall average level. The Shewhart chart detects large deviations (1·5 to 2·0 standard deviations or greater) from the previous stable pattern very quickly, but is not as effective in detecting smaller shifts. This chart considers only the last plotted individual point, thus ignoring information about the process in previous observations.

The moving average control chart is similar in application to the Shewhart chart, but is more effective in detecting small process shifts. For this control chart the moving average of span w, the average of the last w points, is plotted rather than the individual point.

The EWMA control chart, first presented as the 'geometric moving average chart', combines historical data to give less weight to data as they get older:

$$\hat{y}_t = \lambda y_t + (1 - \lambda)\hat{y}_{t-1}$$

where $\hat{y}_t$ is the statistic at time t (the new EWMA), $\hat{y}_{t-1}$ is the statistic at time $t-1$ (the old EWMA), λ is the EWMA weighting parameter and y_t is the observed value at time t (the new observation).[52] The EWMA control chart can be designed to resemble the performance of the Shewhart control chart by the selection of λ in the interval (0, 1). The greater the value of λ, the smaller is the influence of the data in the more distant past. The value of λ can be determined subjectively,[53] but is commonly set at approximately 0·2 in EWMA control chart applications. Since the EWMA is a weighted average of observations, it is less sensitive to the normality assumption and, therefore, provides more flexibility in its application to monitoring problems.

For the 1990 analysis phase, we implemented the SPC charts to monitor and identify aberrations in the forecast errors for 1989 data. For the phase two analysis, we implemented the monitoring system on the 71 weeks of forecast errors available for 1994–1995.

RESULTS

Box–Jenkins Time Series Models

Most of the national and state disease series of reported cases were non-stationary in the mean and/or variance, so required data transformations and/or differencing to attempt to achieve those

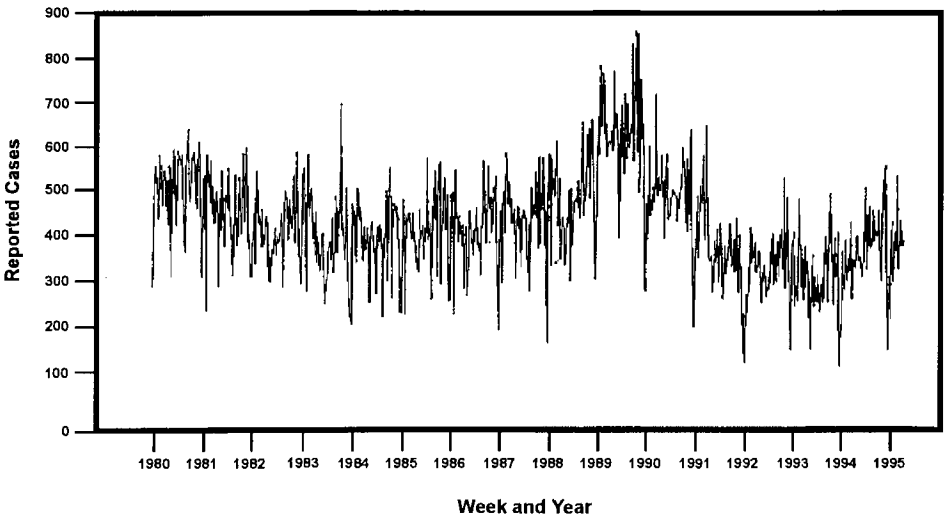


Figure 1. Number of reported cases of hepatitis A, by week, United States, 1980–1995 (provisional data 5 January 1980–13 May 1995)

stationarity requirements for applying Box–Jenkins time series methods. When series with non-stationary means were encountered, we employed differencing and, when non-stationary variances were observed, we used logarithm and square root transformations to increase stability and, hopefully, meet stationarity assumptions. Of the 17 national disease report series, we were able to meet stationarity assumptions and develop adequate Box–Jenkins ARIMA models for hepatitis type A, hepatitis type B, hepatitis non-A and non-B, legionellosis, malaria, meningococcal infections and tuberculosis. Of the state series, we successfully achieved stationarity and developed ARIMA models for all three state series for hepatitis type A and for one state for typhus fever (tickborne). Because of the modelling success at the national and state level for hepatitis type A, we focused analyses on those series and, subsequently, focus reporting here on hepatitis A, although, when meaningful, we make reference to results for other series.

For the hepatitis type A national disease report series (Figure 1), we performed first differencing of order one and 52 to achieve stationarity in the mean for the 1980–1988 data (Figure 2 depicts the first differences of order one) to appropriately apply Box–Jenkins methods for forecasting disease reports for 1989. The final model selected for this series, based on evaluation of statistical fit and model forecasting capabilities, was

$$(1 - 0{\cdot}22B^{52})(1 - B)(1 - B^{52})z_t = (1 - 0{\cdot}90B)(1 - 0{\cdot}82B^{52})a_t.$$

The plot of observed versus model values (Figure 3) indicates the model provides an excellent fit to the data. This model form was the same as and the model coefficients were extremely similar to those of the model developed for forecasting 1988 disease report data based on the 1980–1987 data:

$$(1 - 0{\cdot}21B^{52})(1 - B)(1 - B^{52})z_t = (1 - 0{\cdot}92B)(1 - 0{\cdot}78B^{52})a_t.$$

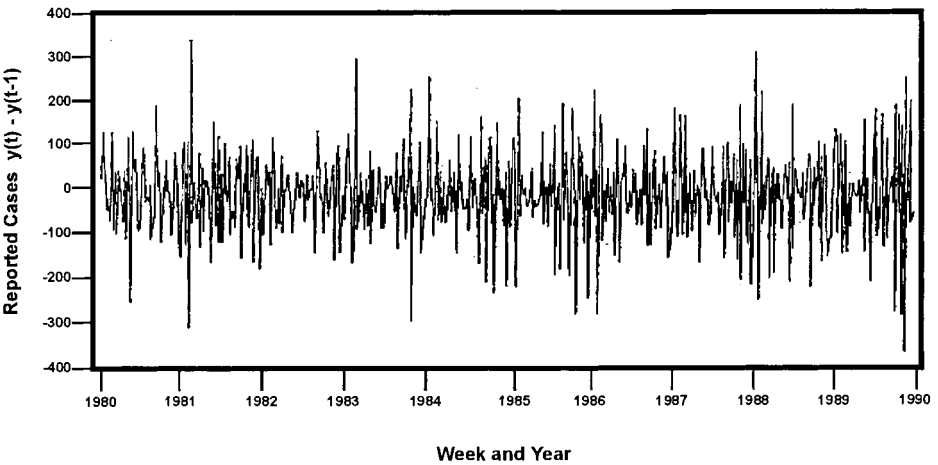


Figure 2. First differences in the number of reported cases of hepatitis A, United States, 1980–1989 (the first difference represents the week-to-week changes in the reported number of cases, $y(t)$)

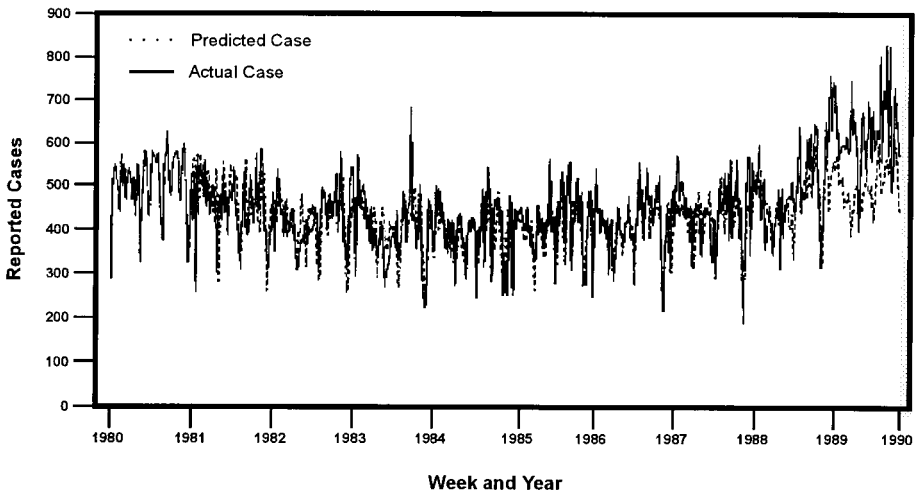


Figure 3. Predicted and actual number of reported cases of hepatitis A, by week, United States, 1980–1989 (the 1981–1988 predicted numbers are model estimates, the 1989 predicted numbers are forecasts and the actual number of cases are provisional data from 5 January 1980 to 30 December 1989)

The final model for hepatitis A for 1980–1993 national data, used to develop the monitoring system for 1994–1995 disease report data, was estimated using the square root of the reported cases data:

$$(1 - B)(1 - B^{52})\, z_t = (1 - 0{\cdot}85B)(1 - 0{\cdot}79B^{52})\, a_t.$$

This model was used to forecast weekly reports for 1994 and the first 19 weeks of 1995 in the two ways given above: (i) forecasted weekly values for all 71 weeks of 1994–1995 in the analysis with

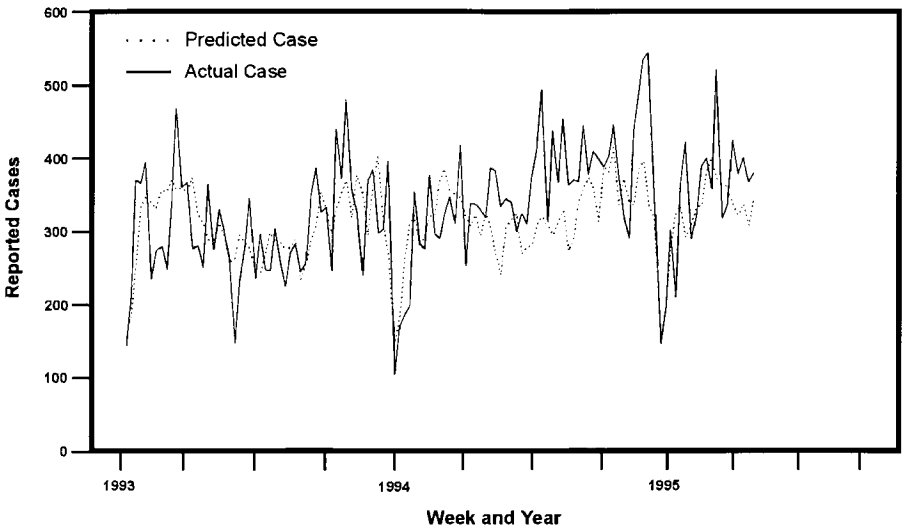


Figure 4. Predicted and actual number of reported cases of hepatitis A, by week, United States, 1993–1995 (the 1993 predicted numbers are model estimates, the 1994–1995 predicted numbers are 71-week-ahead forecasts and the actual number of cases are provisional data from 9 January 1993 to 13 May 1995)

no incorporation of 1994–1995 disease reports or redevelopment of the model form or re-estimation of model coefficients after each weekly forecast was calculated (Figure 4) and (ii) one-week-ahead, rolling forecasts for each of the 71 weeks using the same model form but re-estimating the model coefficients after the addition of each week's newly reported cases from 1994–1995 (Figure 5).

Although we modelled the hepatitis A disease report series for all three states on 1980–1988 data, the forms of the models were different from each other and from that for the national series discussed above.

Statistical Process Control Methods

Shewhart and moving average (with span of size two) control charts developed on national hepatitis type A values forecasted for 1989 (phase one analysis) identified several statistically high values in the data (Figures 6 and 7). The EWMA chart detected a statistically significant deviation from the forecasted or expected values, thus indicating the likelihood of an increasing trend throughout the series (Figure 8).

Of the other six national disease report series for which we developed models and applied the SPC charts in the phase one analysis (through 1989), only for hepatitis B did more than one chart identify any significantly high aberrations (Shewhart and moving average charts identified a few high values close in time and the EWMA chart identified statistical aberrations indicating the likelihood of an increasing trend). For malaria the moving average chart identified a single statistically high aberration. The monitoring system detected no significantly high values for any of the hepatitis non-A and non-B, legionellosis, meningococcal infections and tuberculosis forecasted data. In phase two of the analysis (to week 19 of 1995), none of the three SPC charts

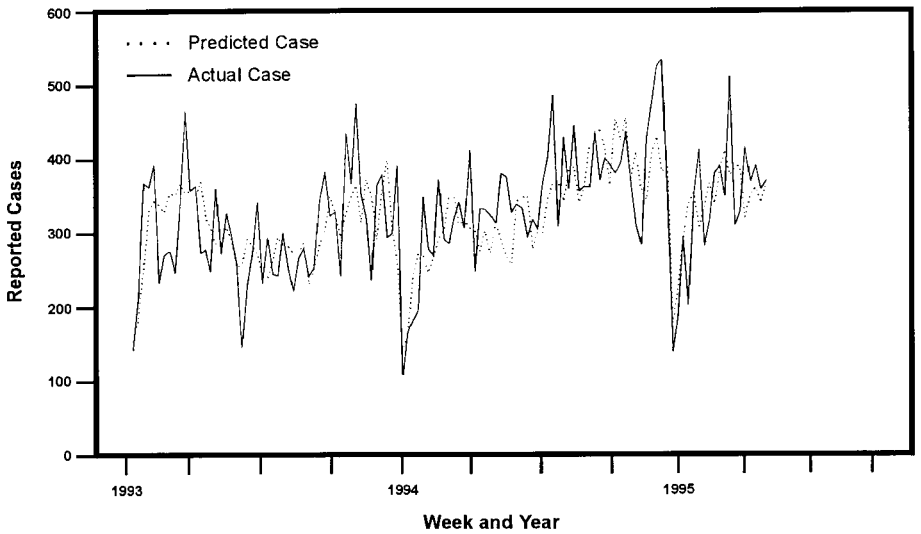


Figure 5. Predicted and actual number of reported cases of hepatitis A, by week, United States, 1993–1995 (the 1993 predicted numbers are model estimates, the 1994–1995 predicted numbers are rolling forecasts and the actual number of cases are provisional data from 9 January 1993 to 13 May 1995)

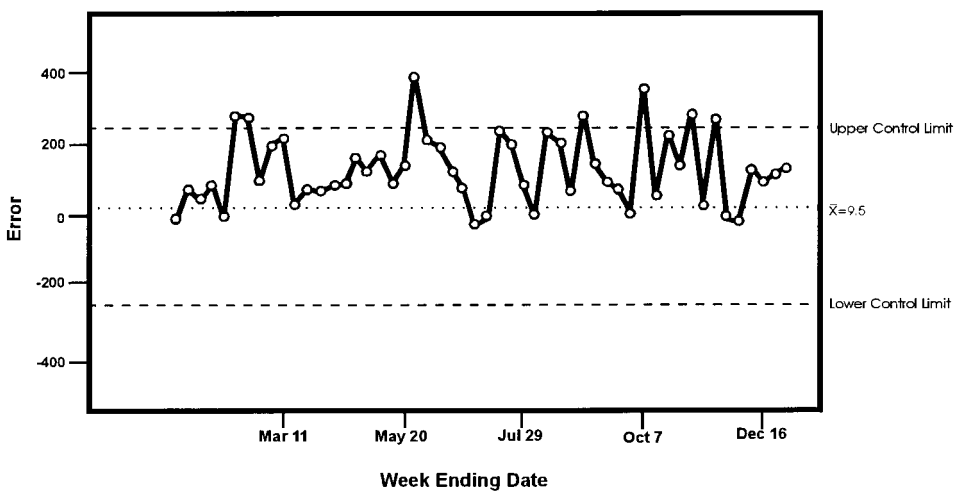


Figure 6. Shewhart control chart for the number of reported cases of hepatitis A, by week, United States, 1989

identified any significantly high values for the 1994–1995 forecasted national hepatitis A disease report series.

For 1989 hepatitis A forecasted data, monitoring results varied by state with no evident relationships for statistically high values among the three states or with the national disease monitoring results. High aberrations were detected by at least one control chart for each of the

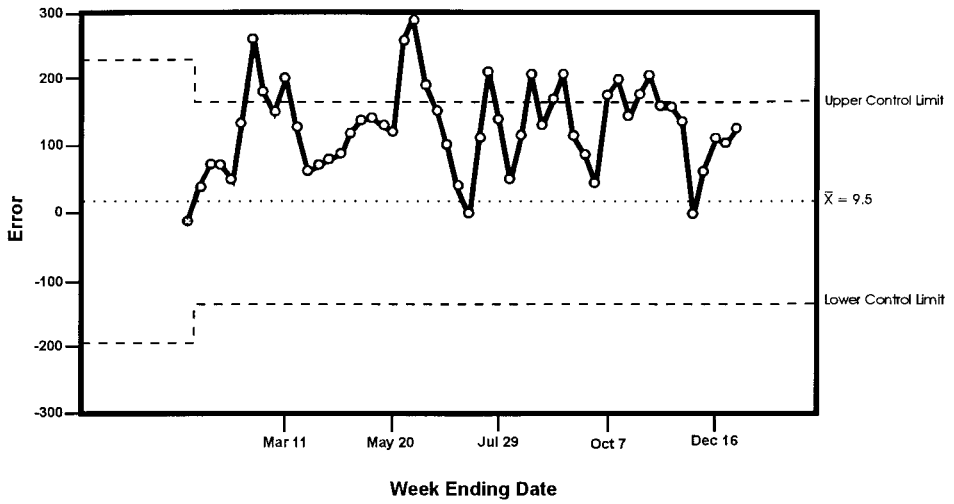


Figure 7. Moving average control chart for the number of reported cases of hepatitis A, by week, United States, 1989

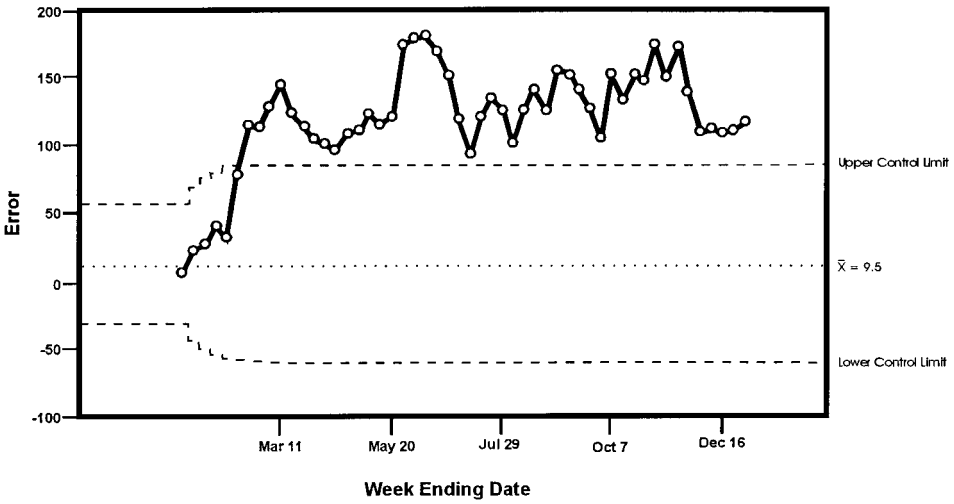


Figure 8. Exponentially weighted moving average control chart for the number of reported cases of hepatitis A, by week, United States, 1989

three states, with all three charts (Shewhart, moving average and EWMA) signalling high values for one state.

DISCUSSION

A critical aspect of public health is to monitor incidence, prevalence and patterns of occurrence and spread of diseases and other health events, and to report those to the public and the public

health community so that appropriate actions are taken. In part because communication technology continues to improve and funding sources require increased justification for scientific research, it is incumbent for public health practitioners to use all available data and analysis tools to make timely, appropriate decisions. We introduce here a monitoring system which holds promise for aiding in identification of unusual observations in public health surveillance data by providing early warning signals.

Because our research focus was on viability of quantitative tools rather than on providing in-depth insight into aetiology or risk factors of disease, we limit the discussion here to use and understanding of the methods, and provide no programmatic guidance regarding disease occurrence, prevention and control. However, the scope and diversity of diseases within the NNDSS provides a substantive arena for evaluation of the monitoring system and has guided us in establishing an ambitious research plan for the future.

We applied our analyses to reported number of cases of disease rather than to reported rates of disease in part because the NNDSS does not include population data, a characteristic shared by most public health surveillance systems. Additionally, because we were investigating potential changes in patterns of disease reports at different times for the same reporting areas, we would not expect dramatic changes in population size for the areas and, thus, analyses of rates would not provide substantial differences in results. Should the focus of analyses shift to comparison among reporting areas with different population sizes, then analyses of rates would be more appropriate. Our methods, however, could readily be applied to rates.

We successfully developed Box–Jenkins ARIMA models for seven of the 17 national disease report series and four of the 17 state disease series we investigated. For those series we were unable to model, the deviations were too great to achieve stationarity and, in some cases, data were too sparse for much of the reporting period to apply the Box–Jenkins methods. This was especially true for the state series when there was a substantial increase in amounts of missing data and low or no reported cases of disease. Another hindrance to the modelling process was the lack of pattern in the disease report series, such as that associated with clustering of reports of disease occurrence at random, non-seasonal times.

A strength of these modelling techniques was the stability demonstrated as additional data became available over two years; recall the form and coefficients of the model were very similar for the 1988 and 1989 forecasting models for hepatitis A. Models developed for the other six national disease series demonstrated this same similarity and stability. With addition of 1990–1993 national hepatitis A data in the phase two analysis, we found the original model form (based on 1980–1989 data) still produced good forecasts, but that a model of substantially different form provided better performance, thus indicating the need to periodically review and change the model form to maintain the best (or perhaps valid) forecasting model and maintain a reliable monitoring system for detecting aberrations.

The greatest utility for Box–Jenkins models comes with one-step-ahead forecasts because forecast errors for longer lead times will in general be correlated. A major advancement in our system occurred when we implemented this one-step-ahead, rolling forecast method in the phase two analysis. Figures 4 and 5 allow comparison of the two forecast methods and show how, when one- week-ahead rolling forecasts are utilized (Figure 5), the forecasts for 1994–1995 more closely mirror the observed values. However, because the model is readjusted with the addition of each new observation when using rolling forecasts, the monitoring system (which would then be implemented on forecasted errors from a continually readjusting model) may have a tendency to miss picking up trends in the disease series, an issue which needs to be evaluated in future work.

A combination or blend of forecasting strategies with specific decision rules may provide the most effective monitoring system in practice.

Because the national disease series is comprised of reports from 50 states, it was not surprising that the forms of the models developed for hepatitis A for the three states were different from the form for the national model. Also, because the three states used in the hepatitis A analysis were geographically dispersed, it was not unexpected that model forms for the states also differed from each other. Future research should include analysis on data from contiguous states to determine whether similar models are developed for those states with similar disease report patterns.

The Shewhart and moving average control charts, both effective at identifying relatively large shifts from the overall average level, signalled high values at similar times for the 1989 hepatitis A forecasted data (Figures 6 and 7). One of the strengths of our monitoring system is employing several SPC charts to detect different types and sizes of aberrations (for example, the EWMA chart can detect small shifts and gradual trends in the data and, thus, complements the capabilities of Shewhart and moving average charts to detect larger, acute changes in the mean of the series). The cumulative sum, or CUSUM, SPC chart is designed to identify smaller sustained shifts in a process and, thus, should yield analysis results similar to those when employing the EWMA chart.[54] We have performed initial work incorporating the CUSUM chart in the monitoring system but, because the research is preliminary, have excluded those results here. VanBrackle and Williamson, in other work presented in this issue, demonstrate how the monitoring system with several control charts detects spike, step and trend changes of varying sizes in the disease series.[36] Their research also includes how correlated data affect capabilities of the monitoring system to detect aberrations. Some research has suggested the use of a combined Shewhart–EWMA control procedure, which should be effective against both large and small shifts.[44] In this combined approach, both the EWMA statistic and the individual forecast error could be plotted on the same chart with the two sets of control limits. However, there is still much work to be done in application of SPC charts within our monitoring system and evaluation of subsequent results, including further application of the CUSUM chart.

There are other important directions related to the problem of detecting aberrations which need to be investigated. Additional methods, such as neural network techniques and fractional differencing, smoothness priors and time-space time series methods, should be applied to the NNDSS data, evaluated, and compared with results from our monitoring system work. Monteiro *et al.* present a comparison of dynamic linear models and ARIMA models on NNDSS data.[41] The effects of multiple testing should be evaluated and perhaps accounted for in the monitoring system. Further work should be developed to examine earliest detection of changes in patterns of disease report data (to identify beginning stages of an epidemic), such as with change point models.[41]

A critical aspect to evaluation of any method is to validate its usefulness in the intended setting. Although we had periodic discussions with disease experts during this research, we need to perform an epidemiologic validation to determine sensitivity and specificity of our monitoring system, realizing that no monitoring system will be able to discern between a disease-related increase and one caused by clerical error, batch reporting or statistical anomaly. This evaluation should be done on the local or state level, from where the data originate and public health responses to substantial increases in reported cases of disease are likely to come.

To affect utilization of analytic methods in rapid response situations, as with potential epidemics, it is crucial that public health practitioners know and understand those methods and have an easy, timely, inexpensive way to implement them. CDC's Statistical Software for Public

Health Surveillance, which provides several methods for analysing surveillance data, including the Box–Jenkins time series techniques, is available through the Internet, but at this time lacks the SPC charts needed to develop the monitoring system described here.[55]

Although this research has focused on our monitoring system for detecting aberrations, it has reinforced thoughts that, because of the breadth and diversity of diseases and the factors affecting them, there is no single method which can be universally applied to public health surveillance data to identify an unusually high number of cases of disease or health event. The monitoring system is a new approach which maximizes strength from two time series methods, and it is an automated, flexible one which shows promise of assisting the public health community in these efforts. However, there remain numerous opportunities for development, application and evaluation of quantitative methods to aid in identifying outbreaks, sentinel public health events and aberrations in disease data, and, thus, facilitate timely actions to decrease unnecessary morbidity and mortality.

REFERENCES

1. Thacker, S. B. and Berkelman, R. L. 'Public health surveillance in the United States', *Epidemiologic Review*, **10**, 164–190 (1988).
2. Teutsch, S. M. and Churchill, R. E. (eds). *Principles and Practice of Public Health Surveillance*, Oxford University Press, New York, 1994.
3. Thacker, S. B., Berkelman, R. L. and Stroup, D. F. 'The science of public health surveillance', *Journal of Public Health Policy*, **10**, 187–203 (1989).
4. Noah, N. D. 'Cyclical patterns and predictability in infection', *Epidemiologic Infections*, **102**, 175–190 (1989).
5. 'National conference on clustering of health events', *American Journal of Epidemiology* (*Supplement*), **132**, S1–S200 (1990).
6. Farr, W. *Progress of epidemics. Second Report of the Registrar General of England and Wales*, His Majesty's Stationery Office, London, 1840, pp. 91–98.
7. Greenwood, M. 'On the statistical measure of infectiveness', *Journal of Hygiene*, **31**, 336–351 (1931).
8. Abbey, J. 'An examination of the Reed-Frost theory of epidemics', *Human Biology*, **24**, 201–203 (1952).
9. Serfling, R. E. 'Methods for current statistical analysis of excess pneumonia-influenza deaths', *Public Health Reports*, **78**, 494–506 (1963).
10. Lui, K.-J. and Kendal, A. P. 'Impact of influenza epidemics on mortality in the United States from October 1972 to May 1985', *American Journal of Public Health*, **77**, 712–716 (1987).
11. Brookmeyer, R. and Liao, J. 'Statistical modelling of the AIDS epidemic for forecasting health care needs', *Biometrics*, **46**, 1151–1163 (1990).
12. Cliff, A. D., Haggett, P. and Stroup, D. F. 'The geographic structure of measles epidemics in the northeastern United States', *American Journal of Epidemiology*, **136**, 592–602 (1992).
13. Cliff, A. D. and Haggett, P. 'Statistical modelling of measles and influenza outbreaks', *Statistical Methods in Medical Research*, **2**, 43–73 (1993).
14. Sprenger, M. J. W., Mulder, P. G. H., Beyer, W. E. P., Strik, R. V. and Masurel, N. 'Impact of influenza on mortality in relation to age and underlying disease, 1967–1989', *International Journal of Epidemiology*, **22**, 334–340 (1993).
15. Cliff, A. D., Haggett, P., Smallman-Raynor, M. R., Stroup, D. F. and Williamson, G. D. 'The application of multidimensional scaling to epidemiologic data', *Statistical Methods in Medical Research*, **4**, 102–123 (1995).
16. Simonsen, L., Clarke, M. J., Stroup, D. F., Williamson, G. D., Arden, N. H. and Cox, N. J. 'A method for timely assessment of influenza-associated mortality in the United States', *Epidemiology*, **8**, 390–395 (1997).
17. Simonsen, L., Clarke, M. J., Williamson, G. D., Stroup, D. F., Arden, N. H. and Schonberger, L. B. 'The impact of influenza epidemics on mortality: introducing a severity index', *American Journal of Public Health*, **87**, 1944–1950 (1997).

18. Ingram, D. D. and Williamson, G. D. 'Statistical and epidemiologic techniques. Summary of major tests available for statistical assessment of clustering, *in* Recommendations and Reports, July 27, 1990', *Morbidity and Mortality Weekly Report*, **39**, (RR-11), 17–23 (1990).
19. Besag, J. and Newell, J. 'The detection of clusters in rare diseases', *Journal of the Royal Statistical Society, Series A*, **154**, 143–155 (1991).
20. Marshall, R. J. 'Mapping disease and mortality rates using empirical Bayes estimators', *Applied Statistician*, **40**, 283–294 (1991).
21. Marshall, R. J. 'A review of methods for the statistical analysis of spatial patterns of disease', *Journal of the Royal Statistical Society, Series A*, **154**, 421–441 (1991).
22. Cressie, N. 'Smoothing regional maps using empirical Bayes predictions', *Geographical Analysis*, **24**, 75–95 (1992).
23. Elliott, P., Cuzick, J., English, D. and Stern, R. (eds). *Geographical and Environmental Epidemiology Methods for Small-Area Studies*, Oxford University Press, New York, 1992.
24. Box, G. E. P. and Jenkins, G. M. *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, 1976.
25. Thacker, S. B. 'Historical development', *in* Teutsch, S. M. and Churchill, R. E. (eds), *Principles and Practice of Public Health Surveillance*, Oxford University Press, New York, 1994.
26. Cates, Jr., W. and Williamson, G. D. 'Descriptive epidemiology: Analyzing and interpreting surveillance data', *in* Teutsch, S. M. and Churchill, R. E. (eds), *Principles and Practice of Public Health Surveillance*, Oxford University Press, New York, 1994.
27. Stroup, D. F., Williamson, G. D., Herndon, J. L. and Karon, J. M. 'Detection of aberrations in the occurrence of notifiable diseases surveillance data', *Statistics in Medicine*, **8**, 323–329 (1989).
28. Smith, A. F. M. and West, M. 'Monitoring renal transplants: an application of the multiprocess Kalman filter', *Biometrics*, **39**, 867–878 (1983).
29. Centers for Disease Control and Prevention. 'Proposed changes in format for presentation of notifiable disease report data', *Morbidity and Mortality Weekly Report*, **38**, (47), 805–809 (1989).
30. Gordon, K. and Smith, A. F. M. 'Modeling and monitoring biomedical time series', *Journal of the American Statistical Association*, **85**, 328–337 (1990).
31. Stroup, D. F. and Thacker, S. B. 'A Bayesian approach to the detection of aberrations in public health surveillance data', *Epidemiology*, **4**, 435–443 (1993).
32. Stroup, D. F., Wharton, M., Kafadar, K. and Dean, A. G. 'Evaluation of a method for detecting aberrations in public health surveillance data', *American Journal of Epidemiology*, **137**, 373–380 (1993).
33. Wharton, M., Price, W., Hoesly, F., Woolard, D., White, K., Greene, C. and McNabb, S. 'Evaluation of a method for detecting outbreaks of diseases in six states', *American Journal of Preventive Medicine*, **9**, 45–49 (1993).
34. Nobre, F. F. and Stroup, D. F. 'A monitoring system to detect changes in public health surveillance data', *International Journal of Epidemiology*, **23**, 408–418 (1994).
35. Stroup, D. F. 'Special analytic issues', *in* Teutsch, S. M. and Churchill, R. E. (eds), *Principles and Practice of Public Health Surveillance*, Oxford University Press, New York, 1994.
36. VanBrackle, L. and Williamson, G. D. 'A study of the average run length characteristics of the National Notifiable Diseases Surveillance System', *Statistics in Medicine*, **18**, 3309–3319 (1999).
37. Choi, K. and Thacker, S. B. 'An evaluation of influenza mortality surveillance, 1962–1979.I. Time series forecasts of expected pneumonia and influenza deaths', *American Journal of Epidemiology*, **113**, 215–226 (1981).
38. Helfenstein, U. 'Box-Jenkins modelling of some viral infectious diseases', *Statistics in Medicine*, **5**, 37–47 (1986).
39. Stroup, D. F., Thacker, S. B. and Herndon, J. L. 'Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age, 1962–1983', *Statistics in Medicine*, **7**, 1045–1059 (1988).
40. Kopjar, B. and Guldvog, B. 'Time variations in injury incidence', *National Institute of Public Health Annals*, **16**, 3–10 (1993).
41. Monteiro, A. B. S., Telles, P. R., Nobre, F. F. and Williamson, G. D. 'A comparison of epidemiological time series forecast methods: Bayesian and ARIMA models' (unpublished manuscript).
42. Frisen, M. 'Evaluations of methods for statistical surveillance', *Statistics in Medicine*, **11**, 1489–1502 (1992).

43. Centers for Disease Control and Prevention. *Summary of Notifiable Diseases, United States 1996, Morbidity and Mortality Weekly Report*, **45**(53), 1997.
44. Montgomery, D. C. *Introduction to Statistical Quality Control*. John Wiley & Sons, New York, 1991.
45. Yourstone, S. A. and Montgomery, D. C. 'A time-series approach to discrete real-time process quality control', *Quality and Reliability Engineering International*, **5**, 309–317 (1989).
46. Montgomery, D. C. and Mastrangelo, C. M. 'Some statistical process control methods for autocorrelated data (with discussion)', *Journal of Quality Technology*, **23**, 179–204 (1991).
47. Makridakis, S., Wheelwright, S. C. and McGee, V. E. *Forecasting: Methods and Applications*. John Wiley & Sons, New York, 1983.
48. Pankratz, A. *Forecasting with Univariate Box–Jenkins Models Concepts and Cases*. John Wiley & Sons, New York, 1983.
49. *SAS/ETS*^{R} *User's Guide*. SAS Institute Inc., Cary, NC, 1993.
50. Shewhart, W. A. *Economic Control of Quality of Manufactured Product*, Van Nostrand, New York, 1931.
51. *SAS/QC*^{R} *Software: Reference*. SAS Institute Inc., Cary, NC, 1990.
52. Roberts, S. W. 'Control chart tests based on geometric moving averages', *Technometrics*, **1**, 239–250 (1959).
53. Hunter, J. S. 'The exponentially weighted moving average', *Journal of Quality Technology*, **18**, 203–210 (1986).
54. Page, E. 'Continuous inspection schemes', *Biometrika*, **41**, 100–115 (1954).
55. Haddad, S. F., Dean, A. G., Williamson, G. D. and Stroup, D. F. *Statistical Software for Public Health Surveillance*, Centers for Disease Control and Prevention, Atlanta, 1994.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3299–3307 (1999)

THE CUMULATIVE q INTERVAL CURVE AS A STARTING POINT IN DISEASE CLUSTER INVESTIGATION

RINA CHEN*

Department of Applied Mathematics, Israel Institute for Biological Research, P.O.Box 19, Ness-Ziona, Israel, 74100

SUMMARY

Statistical analyses aimed at detection and investigation of clustering are associated with inherent difficulties. Both types of statistical errors are large in these analyses. The results of the analyses should indicate whether or not at least some of the cases are clustered, and if they are, whether or not the cluster is related to an exposure. The temporal changes in the incidence rate of the disease may alleviate the difficulties associated with the large statistical errors. Because of the sparse data, estimates of the incidence rates over time are not reliable. In this study we present the q interval statistic that has the uniform (0,1) distribution. It can be viewed as a standardized time interval between consecutive diagnoses of the disease. As such, it reflects the reciprocal of the incidence rates. Since it is measured for each diagnosis, it is sensitive to gradual change in the incidence rate, and in general to a true clustering that is due to exposure, even when the test result is not significant. When clustering is detected, it may indicate which of the possible reasons leading to a cluster has a sound basis. As a result, the epidemiological search for exposure is limited to situations indicated by the q intervals. In addition, the q interval presents a useful survival statistic in a follow-up study when no control group is available. Software programs in SAS and in SYSTAT are available. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Health professionals are often required to conduct an investigation regarding an apparent cluster of cancer cases in a community. The requirement may stem from alarm elicited among the community members, or from statistical analyses of data observed in that or in a related community. Investigation aimed at identification of the cause of the alarming cluster is then required. Several inherent difficulties are involved in the statistical methods designed for this investigation, as well as for detection of clustering of rare events.[1] These include large type I statistical error as well as low power. The type I error is inflated because of *ad hoc* analyses. The low efficiency of the methods is related to the fact that the data are sparse, the incubation period is long and the impact of each carcinogen is minor. In some settings the efficiency of the analysis is further attenuated because the time of exposure is unknown and/or because of the 'healthy worker effect'.[2] These difficulties may be somewhat alleviated if one considers the temporal pattern of the data as related to the possible situations associated with the disease clustering. In this study we present a statistic that is sensitive to the temporal changes in the incidence rate, and show its value in the detection and interpretation of clusters of a rare disease.

* Correspondence to: Rina Chen, 37, Tumpeldor St., Tel-Aviv 63425, Israel.

CCC 0277–6715/99/233299–09$17.50

2. METHODS

In general, a cluster may be related to either one of the following situations: the community has been exposed to a local carcinogen, realization of a rare chance has occurred, or the disease is being diagnosed at an earlier stage than before (because of a new medical device, for example).

The temporal changes in the incidence rate of the disease may indicate which of the three possibilities is a likely explanation. These changes may reflect on the number and temporal position of clusters. When the alarm is caused by exposure to a local carcinogen, the data are expected to show a single cluster. This cluster is likely to include the later diagnoses when the exposure is relatively new, and all the diagnoses when it is old. When the clustering is an incidental event, the data may have more than just one cluster of cases. When clustering is due to enhanced diagnoses, a single cluster which may occupy the midst of the data set is expected. This clustering may be followed by an incidence rate that is even lower than that expected. Thus, the temporal pattern of the incidence rate may indicate which of the possible explanations to the alarm has a sound basis. This pattern may be indicative even when the results are not significant. Estimates of these rates within the study period are usually not stable, due to the fact that the cases are few and spread over a long period. However, the temporal pattern of the time interval between consecutive diagnoses may be indicative. This time interval is the waiting time until the next diagnosis, as such it reflects the inverse of the incidence rate within that interval, and can therefore feature the temporal pattern of the incidence rate

The q interval statistic was recently suggested for this purpose[3] and applied to registered data of several leukaemias in a city in Israel.[4] This statistic is defined as the null *a priori* probability that no diagnosis is observed within the given time interval. It is a random variable of the uniform (0,1) distribution.[5] It should be noted that although q is expressed as a probability, it is a random variable since the associated time interval (that is, the observed interval between two diagnoses) is a random variable. The fact that the distribution of q is known and that it is a standardized time interval (as shown below), in the sense that changes in the size and profile of the population are controlled, renders it useful in statistical tests and in graphical display of the temporal changes of the incidence rate.

The probability that no diagnosis is observed within a given time interval can be evaluated using the specific (for example, for age and gender) incidence rates in a reference population. For this evaluation we first calculate the null expected number of diagnoses for each year in the observed interval as the sum $\sum r_i n_i$, where r_i and n_i, respectively, are the incidence rate and the size of stratum i in that year. Then, the expected number of diagnoses are summed over the years included in the observed interval to get the expected number of diagnoses within that interval, denoted by L. For example, suppose that the dates of diagnosis of the $i-1$ and the i cases are 1 January 1980 and 15 November 1984, respectively, and the expected number of diagnoses in each year during 1980–1984 is

Year	Number expected
1980	0·308
1981	0·312
1982	0·348
1983	0·379
1984	0·411

The expected number of diagnoses between 1 January 1980 and 15 November 1984 is

$$L = 0{\cdot}308 + 0{\cdot}312 + 0{\cdot}348 + 0{\cdot}379 + \left(\frac{10{\cdot}5}{12}\right) 0{\cdot}411 = 1{\cdot}707.$$

We now define a standard time unit as the expected calendar time until the next diagnosis. The number of months in each unit may change from one diagnosis to the next, depending on changes in the size and structure of the population. Under this definition, L is the number of units in a given interval in terms of the expected number of diagnoses. Since the number of diagnoses in a given period is the sum of a time dependent Poisson process,[6,7] the distribution of the time interval q between diagnoses is exponential. Thus

$$q = \exp(-L). \tag{1}$$

In our example

$$q = \exp(-1{\cdot}707) = 0{\cdot}181.$$

Here we define a time unit that is tailored to the conditions prevailing during each year. Thus, q is a standardized statistic in the sense that differences between communities (or even diseases) are controlled by tailoring the definition of the time unit according to the null expected number of diagnoses. Temporal changes in the size and structure of the community are controlled by updating the time unit definition.

When the incidence rate is elevated to γ times the baseline rate, the assumed length of the time interval is L but the actual length is L/γ units, since the actual expected time between diagnoses is $1/\gamma$ of that assumed under the baseline rate. The probability that the interval is larger than L is

$$\Pr(x > L) = \exp(-L/\gamma). \tag{2}$$

It can easily be shown that the expected value for q is

$$E(q) = \frac{\gamma}{(\gamma + 1)}. \tag{3}$$

Consecutive q intervals are expected to be larger than 0·5 when $\gamma > 1$, and the slope of the curve gets steeper as γ increases. A curve of the q intervals accumulated over consecutive diagnoses may indicate if and when the incidence rate has risen. A crude estimate of γ can be obtained as the ratio between the observed and the expected number of diagnoses during the elevated rate period. This is only a crude estimate since the relevant period is determined by the data. Another crude estimator of γ presented below is based on the expected median value of q under the elevated rate.

When $\gamma > 1{\cdot}0$, the distribution of the q intervals is no longer uniform. However, a crude estimate of γ can be obtained, using q_{m}, the median value of the q intervals. In this case, $L\gamma$ diagnoses are expected within the median interval but only L are assumed. Thus, we have

$$0{\cdot}5 = \exp(-L\gamma)$$

and

$$q_{\mathrm{m}} = \exp(-L)$$

hence

$$0{\cdot}5 = q_{\mathrm{m}}^{\gamma}$$

and

$$\gamma = \ln(0{\cdot}5)/\ln(q_m). \tag{4}$$

Unlike the ratio (observed/expected number of cases), this estimate is not much affected by the large intervals associated with incidental clustering. In general, the ratio provides a more stable estimate than the median, but q_m provides a better estimate when it is difficult to determine the period of high incidence. Because of the long and variable latency, this period may not be cleared even when the clustering is related to a new carcinogen.

Computer programs in SAS (by G. Haugh of ATSDR, Atlanta) and in SYSTAT that calculate q, cumulative q, and present curves of the observed and the expected cumulative q intervals, are available upon request. The use of these programs requires two files: one to provide the expected number of diagnoses at each year covered by the study, and one to provide the date (at least month and year) of diagnosis for each case.

3. EXAMPLES

Using the q interval we were able to interpret and detect clusters in real data sets. These include interpretation of clustering of brain tumour in a residential population as caused by enhanced diagnoses (Chen and Kaye, unpublished) and detection of clustering of colon cancer cases among asbestos workers, even though the overall number of cases was lower than expected (due perhaps to the 'healthy worker effect'). In the second instance, the slope of the cumulative q was smaller than expected at the early period and clearly higher than expected later. Here we present the interpretation of three other examples. These include:

1. Clustering of breast, uterine and ovarian (BUO) cancer cases among the 171 individuals of a workplace.
2. Incidental clustering of all cancers, except for the BUO cancers observed in the workplace community of example 1.
3. Clustering of chronic myeloid leukaemia in a city in Israel, based on data observed during 1960–1990.

The first two examples are related to a laboratory workplace in Israel.[1,8] The analyses of these data were made in response to the workers' suspicion that the number of cancer cases among them was excessively large. Since many different chemicals were being used in this laboratory, the possibility that the workers were exposed to several carcinogens which may have triggered several cancers is not an unreasonable assumption. In total there were 16 cases observed versus 7·18 expected (assuming Poisson distribution, $p < 0{\cdot}003$). Most of the cases were females and the investigators suspected that the high incidence rate for all cancers was accounted for by the high rate of the three cancers associated with hormonal activity, that is, breast, uterine and ovarian cancers. The cases were divided into two subsets and the associated q intervals were calculated. Most of the workers were young when they started working in the laboratory, so a delay in the possible effect due to the long incubation period or the 'healthy worker' effect was not expected.

Table I presents the dates of diagnoses and the q intervals for BUO cancers in the workplace population. The q interval of the first case was calculated (using equation (1)) using the number of diagnoses expected between the recruitment date of the first female worker (in 1951) and the diagnosis date of the first case (9/1980). The q interval for the second case in the table is based on the number of BUO diagnoses expected between 9/1980 (the date of the first diagnosis) and

Table I. Diagnosis dates and q intervals of BUO cancer cases

Case number	Date	q
1	9/80	0·545
2	10/80	0·993
3	4/81	0·957
4	1/82	0·933
5	6/85	0·651
6	3/87	0·764
7	11/89	0·600
8	8/90	0·851

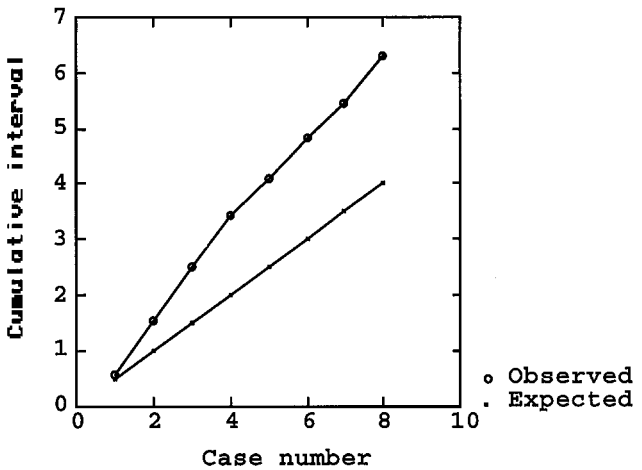


Figure 1. Cumulative q intervals of BUO cancers in a workplace

10/1980 (the date of the second diagnosis). During 1980, 86 women were at risk. This number includes women who joined the staff during or before 1980 and did not die, or had been diagnosed prior to 1980. For each woman at risk, the probability to be diagnosed during 1980 was taken as the (age and ethnic origin) specific incidence rate in Israel at that period and the proportion of time she was at risk during that year. The sum of these probabilities over the 86 individuals at risk is 0·086 and is the expected number of diagnoses during 1980. Since the gap between the two diagnoses is (at most) one month, $L = 0{\cdot}086/12 = 0{\cdot}007$ and $q = \exp(-0{\cdot}007) = 0{\cdot}993$.

In total, there were eight BUO cancer cases as compared to 2·93 expected ($p \leqslant 0{\cdot}01$). The slope of the cumulative q (Figure 1) is consistently higher than that of the expected curve, reflecting the fact that the q intervals are all larger than 0·5. These results strongly indicate that the high incidence rate of BUO cancers is due to exposure to a local carcinogen.

For the entire period γ was estimated to be equal to 2·73 using the ratio of observed and expected number of cases, and 3·24 by the median q. Since the rate seems to be elevated over the entire period, $\gamma = 2{\cdot}73$ was considered a more reliable estimate.

Table II. Diagnosis dates and q intervals of cancer cases other than BUO

Case number	Date	q
1	12/72	0·520
2	2/76	0·705
3	1/80	0·492
4	11/84	0·312
5	1/87	0·537
6	8/89	0·397
7	9/90	0·647
8	12/90	0·903

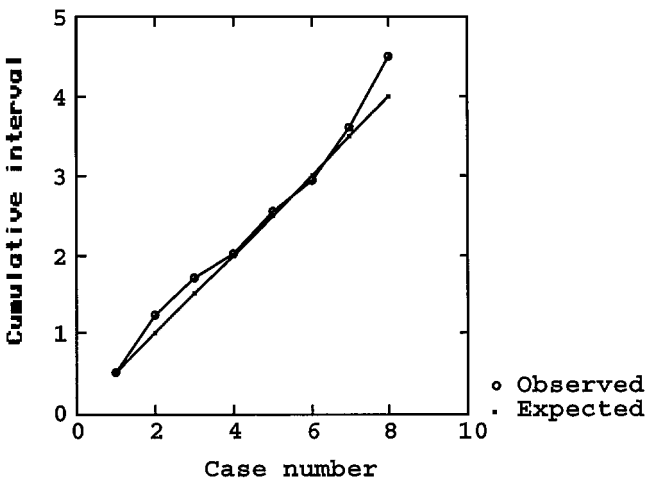


Figure 2. Cumulative q intervals for cancers other than BUO cancers in a workplace

Table II presents the dates of diagnosis and the associated q of all but the BUO cancers diagnosed until December 1990. There were eight cases observed as compared to 4·3 expected ($p \leqslant 0{\cdot}07$). Even though the results are on the verge of significance, the fact that there is no consecutive set of large q values indicates that this cluster may be a rare chance occurrence. Figure 2 presents the observed and expected cumulative q curves. Two clusters with an intervening period appear from this curve. Thus, the relatively large number of cases seems to be a rare chance occurrence rather than one related to a carcinogen. This interpretation is supported by the fact that the crude estimates of γ and of q_m are close to the expected values; γ is 1·86 by the ratio and 1·09 by q_m.

The third example relates to cases of chronic myeloid cancer in a residential population during 1960–1990. The population in this city (Ashkelon, Israel) was about 34,000 in the 1960's and grew to about 56,000 in the 1980's. The baseline rates for these data are based on the gender and age group specific incidence rates of the Jewish population in Israel in each of three periods (1960–1970,1971–1980 and 1981–1990). The product of a stratum incidence rate and the size of

Table III. Diagnosis dates and q intervals of chronic myeloid cancer cases in a city in Israel during 1960–1990.

Case number	Age (sex)	Date	q
1	29(m)	10/66	0·193
2	52(f)	– /67	0·812
3	43(m)	6/72	0·237
4	77(m)	– /77	0·251
5	26(f)	– /77	0·884
6	77(f)	3/77	0·884
7	4(f)	– /78	0·749
8	67(f)	5/80	0·542
9	76(f)	7/84	0·246
10	78(f)	– /85	0·708
11	85(m)	11/86	0·607
12	43(f)	4/87	0·859
13	60(f)	10/88	0·576

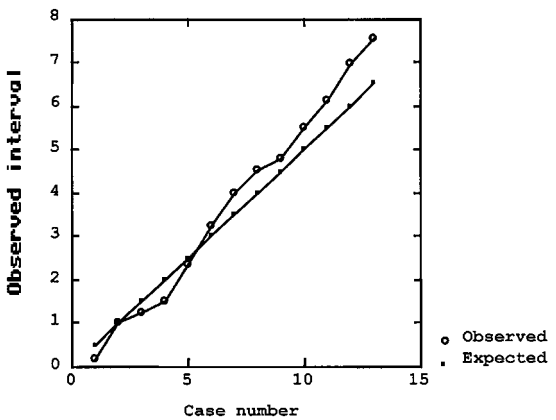


Figure 3. Cumulative q intervals of chronic myeloid cancer cases in a city in Israel during 1960–1990

the corresponding population in Ashkelon in year t is the expected number of diagnoses for that stratum during that year. The expected number of diagnoses during year t is the sum of these expected numbers over the strata. Using (1) where L is the number of diagnoses expected between the two diagnoses, we get the q interval.

Table III presents the dates and q intervals of the data. Judging from the q intervals and the slope of the curve relative to the expected curve, it was concluded that the rate of the disease was close to the baseline rate until 1977, and then was elevated (Figure 3). From the point of elevation on, only one of the nine q intervals is below 0·5. Assuming that the elevation point of time is in mid-1977, the estimate of γ is 1·90 by the observed/expected ratio and 2·0 by q_{m}.

4. DISCUSSION

The analysis and interpretation of data associated with temporal clustering of cancer (or of any other rare health event) are problematic because the statistical errors of both types are large. One reason for the low efficiency is that the location of the cluster on the time axis cannot usually be hypothesized, because the exposure and the incubation period are unknown. It is possible that only the later diagnoses in the data set are clustered, whereas the overall number of cases is close to that expected. Our experience showed one instance in which the count was even smaller than expected, but a clear clustering of the last events was clear from the cumulative q curve. This instance was related to colon cancer deaths among asbestos workers in Israel. In this instance, 10 deaths were observed during 1978–1992 as compared to 14·5 expected.

Statistical tests that are sensitive to this situation may be derived from the CUSUM[9], the Scan,[10] the Cuscore[3] and the Sets[3] monitoring techniques. These techniques may be more efficient than a test based on the overall count of cases when the rise in the incidence rate is late (due to incubation period etc.). However, in view of the small number of cases usually involved, the power of these tests may not be adequate. Usually these tests are taken in *ad hoc* situations, therefore when the result is significant, the possibility that the cluster involved is due to other reasons than exposure to a local carcinogen should be seriously considered. In view of the above problematic issues it is clear that, in addition to and above conducting a formal test, the temporal pattern of the diagnoses in the data set may contribute to the detection and the interpretation of temporal clustering whether or not the result of the statistical test is significant. The q and the cumulative q intervals display this pattern.

The q interval is actually a standardized statistic of the calendar time interval. As such, it reflects the reciprocal of the average incidence rate of the specified interval. Since it is a measure associated with each diagnosis, it is sensitive to temporal changes in the incidence rates resulting from exposure, enhanced diagnoses or random occurrence. The consistency over several diagnoses may indicate which of the possible interpretations is a likely explanation.

It is important to note that the accuracy of this interval depends on the accuracy of the baseline rates. Usually these rates are estimates specific to age and gender that are based on a large sample. If these rates are biased, the q intervals may be on the average smaller or larger than 0·5 even though the actual incidence rates equal those of the reference population (that is, the baseline rates). This possibility should be taken into account when interpreting a cumulative q interval curve with a higher than expected slope. When the slope is high over all the cases of the data set, this possibility is more likely than when the slope becomes steeper within the series. Thus, in example 3, the consistently large q values after 1977 seem to reflect a real increase in the incidence rate.

The q interval has wide applicability, as it can also be used in survival analysis of a follow-up study when no control group is available. Such studies play an important role in detecting environmental health hazards in a residential or a workplace community. Using this approach we were able to analyse breast cancer data from the TCE Trichloroethylene subregistry of the National Exposure Registry (NER).[11] For this analysis we used the age specific incidence rates of the SEER data.[12] No clear evidence of clustering was observed (unpublished). In these situations, the results reflect the survival in the follow-up group relative to that expected according to the incidence rates in the reference population. In contrast, the Kaplan–Meier survival curve represents the survival experience of the group itself. This experience is compared with that of a control group in order to evaluate the exposure effect. Both statistics, the q and the

Kaplan–Meier, are estimated at each event. However, whereas the K–M statistic estimates the *observed* survival experience of the follow-up group up to the event, the q statistic is an estimate of the *expected* survival between the last two events. The effect of the exposure on survival is likely to be apparent earlier when using the cumulative q than when using the Kaplan–Meier statistic. This is because the latter statistic is usually close to 1·0 in both the follow-up and control groups, even in the presence of exposure effect. Therefore, the gap between the two groups is likely to be small for the first few events. Consider for example the survival of groups of size 1000 each. Suppose that the first three events occurred in the exposed group. In this case the observed survival is 0·997 in the exposed group and 1·000 in the control. The q interval is expected to be higher than 0·5 with the first diagnosis in the cluster and with each of the latter diagnoses. For example if $\gamma = 1{\cdot}5$, the expected q is 0·60 for each event.

In conclusion, the q interval represents the survival experience of the exposed group relative to that expected in the reference population. The temporal pattern of survival is a useful clue in the epidemiological investigation of an observed cluster, and the difference between the observed and expected cumulative q curves enables an evaluation of the exposure effect in a follow-up study.

REFERENCES

1. Chen, R. 'Exploratory analysis as a sequel to suspected increased rate of cancer in a small residential or workplace community', *Statistics in Medicine*, **15**, 807–816 (1996).
2. Rothman, K. J. *Modern Epidemiology*, Little, Brown and Company, Boston/Toronto, 1986.
3. Chen, R. 'Detection of cancer subtle "epidemics" in small communities', PhD Thesis, Sackler School of Medicine, Tel-Aviv University, 1995.
4. Chen, R., Iscovich, J. and Goldbourt, U. 'Clustering of leukaemia cases in a city in Israel', *Statistics in Medicine*, **16**, 1873–1887 (1997).
5. Hogg, R. V. and Craig A. T. *Introduction to Mathematical Statistics*, 2nd edn, Macmillan Company, 1965.
6. Cox, D. R. and Lewis, P. A. W. *The Statistical Analysis of Series of Events*, Methuen and Co Ltd., London, 1966, p. 28.
7. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1992, pp. 136–137.
8. Modan, B., Blumstein, Z., Luxenburg, O., Novikov, I. and Shemer, J. 'A potential risk of cancer in central laboratory', *Statistics in Medicine*, **15**, 759—763 (1996).
9. Levin, B. and Kline, J. 'The cusum test of homogeneity with an application in spontaneous abortion epidemiology', *Statistics in Medicine*, **4**, 469–488 (1985).
10. Wallenstein, J., Naus, J. and Glaz, J. 'Power of the scan statistic for detection of clustering', *Statistics in Medicine*, **12**, 1829–1843 (1993).
11. Burg, J. R. 'Policies and procedures for establishing the National Exposure Registry', *Journal of the American College of Toxicology*, **8**, 949–954 (1989).
12. Ries, L. A. G., Hankey, B. F., Miller, B. A., Hartman, A. M. and Edwards B. K., *Cancer Statistics Review 1973–1988*, National Cancer Institute. NIH Pub. No. 91-2789, 1991.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3309–3319 (1999)

A STUDY OF THE AVERAGE RUN LENGTH CHARACTERISTICS OF THE NATIONAL NOTIFIABLE DISEASES SURVEILLANCE SYSTEM

LEWIS VANBRACKLE[1]*, AND G. DAVID WILLIAMSON[2]

[1]*Department of Mathematics, Kennesaw State University, 1000 Chastain Road, Kennesaw, GA 30144, U.S.A.*
[2]*Centers for Disease Control and Prevention, 2877 Brandywine Road, MS K-73, Atlanta, GA 30341, U.S.A.*

SUMMARY

This study examines the statistical properties (that is, false positive and negative signals) in detecting unusual patterns of reported cases of diseases from the Centers for Disease Control and Prevention's National Notifiable Diseases Surveillance System. Control charts are applied to the residuals of one-step ahead forecasts based on Box–Jenkins models of reported cases of disease. Simulation and analytical techniques are used to study the average run length characteristics of these control charts for various types of changes in the levels of the series, including spike, trend and step changes. The average run lengths for the highly correlated disease series are much longer than for the usual independent data case. This increase in the average run lengths is strongly influenced by the type of change in the level of the series and by the type of control chart. Understanding the average run length characteristics of the control charts can lead to timely detection of changes in the levels of disease series, and subsequent timely public health actions to decrease unnecessary morbidity and mortality. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

The National Notifiable Diseases Surveillance System (NNDSS) tracks weekly reported cases of 45 infectious diseases at the state and national levels. There is great interest in using the data from the NNDSS to detect significant aberrations, particularly increases, in the reported number of cases of any of the 45 diseases tracked by the system. Seventeen diseases thought to have good modelling potential were chosen from the NNDSS by Williamson and Weatherby.[1] Although state level data would be preferable for practical applications, the national level of reporting was chosen for ease of modelling. Williamson and Weatherby successfully identified and fit seasonal autoregressive integrated moving average (ARIMA) models (Box and Jenkins[2]) for seven of the 17 chosen disease series. The successfully modelled diseases were hepatitis A, hepatitis B, hepatitis non-A–non-B, legionellosis, malaria, meningococcal infections and tuberculosis.

This paper examines the application of statistical process control charts to the residuals from one-step ahead forecasts of ARIMA modelled time series in order to detect significant aberrations

* Correspondence to: Lewis VanBrackle, Department of Mathematics, Kennesaw State University, 1000 Chastain Road, Kennesaw, GA 30144, U.S.A. E-mail: Lvanbrac@kennesaw.edu

CCC 0277–6715/99/233309–11$17.50

in the level of the modelled process (Alwan and Roberts[3]). The average run length (ARL) characteristics are found for four common control charts and four types of changes in the level of the process.

The recursive relationship of the residuals, based on the results of Wardell *et al.*,[4] is used to simulate ARLs. The distribution of the residuals is derived and used in a Markov chain approach similar to that of Lin and Adams[5] to calculate ARLs analytically. The results of the simulation and the analytical methods agree quite well. The results show that the ARLs are substantially influenced by the seasonal behaviour of the models. Generally, the ARLs are much larger for the control charts applied to residuals from the ARIMA models than for the control charts applied to independent data, for which the charts were developed. Those models with the weakest seasonal behaviour have the largest ARLs.

2. CONTROL CHARTS

2.1. The Shewhart Control Chart

The simplest form of the control chart is the Shewhart chart. In a Shewhart chart for monitoring the level of a process, an observation at time t, x_t, is used to indicate whether the process has undergone some shift in its level. A value of $|x_t|$ exceeding certain control limits indicates that the process level has shifted from its previous level. Control limits are usually expressed in terms of the process standard deviation and are chosen to give the chart a good balance between failing to indicate a real shift in process level (a type II error) and indicating a shift when none has occurred (a type I error). An important characteristic of the Shewhart chart is its rapid detection of large shifts in the process level. However, the Shewhart chart is slow to detect small or moderate changes in the process level.

2.2. The moving average control chart

The moving average (MA) control chart uses the moving average of observations of the process as the control statistic and is more sensitive than the Shewhart chart to small shifts in the level of the process. The control statistic of the MA chart with span m is given by

$$y_t = \frac{x_{t-m+1} + x_{t-m+2} + \ldots + x_t}{m}.$$

The chart signals when $|y_t|$ exceeds control limits expressed as a multiple of the standard deviation of y_t. The particular multiple is chosen to give the chart good properties, as discussed for the Shewhart chart. The larger the span of the MA chart, the more sensitive the chart is to small shifts in the process level. In our study, we have used a span of two. Thus our MA chart may be expected to provide somewhat better detection of small shifts in the process level than the Shewhart chart.

2.3. The Exponentially Weighted Moving Average Control Chart

The exponentially weighted moving average (EWMA) control chart (Roberts[6] and Hunter[7]) uses the control statistic $y_t = (1 - \lambda)y_{t-1} + \lambda x_t$ where $0 < \lambda \leqslant 1$ is a parameter of the chart. The chart signals that the process level has changed when $|y_t|$ exceeds control limits expressed as a multiple

of the asymptotic standard deviation of y_t. Again, the particular multiple is chosen to give the chart good properties.

The EWMA chart is sensitive to small shifts in the process level for small values of λ and to large shifts in the process level for large values of λ. Note that for $\lambda = 1$, the EWMA chart is identical to the Shewhart chart. In our study, we have used $\lambda = 0{\cdot}25$. This value of λ is commonly used in industrial applications and is within the range of 0·1 to 0·5 often suggested in the literature for detecting shifts of one-half to one standard deviation of the process.

2.4. The Cumulative Sum Control Chart

The cumulative sum (CUSUM) control chart (Page[8]) is based on sums of observations and can be sensitive to small shifts in the process level. The CUSUM chart uses one statistic for detecting a positive shift in the process level and another statistic for detecting a negative shift in the process level. The statistic for detecting a positive shift is $y_t = \max(0, y_{t-1}) + x_t - h$ where h is a parameter of the chart. This CUSUM statistic accumulates evidence of a positive shift in the level of the process. If there is no evidence of a positive shift, the CUSUM statistic resets to zero. The CUSUM signals that a positive shift in the process level has occurred when y_t exceeds the control limit. As usual, the control limit is chosen to give the CUSUM chart good properties.

A similar procedure using $y_t = \min(0, y_{t-1}) + x_t + h$ and signalling when y_t is less than the control limit is used to detect negative shifts in the process level. To detect both positive and negative shifts in the process level, we have employed the standard procedure of using two one-sided charts concurrently.

The CUSUM procedure can be derived from a sequential probability ratio test. This derivation indicates that the value of h should be chosen as one-half of the shift in the level which should be detected quickly. In our study, we used $h = 0{\cdot}25$, indicating that we want quick detection of a shift of one-half of the standard deviation of the process.

3. THE AVERAGE RUN LENGTH

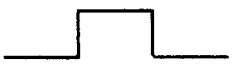
In studying the properties of control charts, the emphasis has been on determining the ARL of the chart. The ARL of a chart is the expected number of samples to be taken (in our case the expected number of weeks) before the chart indicates a shift in the process level. The ARL should be large when there has been no change in the process, but the ARL should be small when the process has undergone a change. Typically ARLs are evaluated for zero shift in the process level (in-control ARLs) and for several shift values which should be detected quickly. After consulting with epidemiologists at the Centers for Disease Control and Prevention, it was decided that in-control ARLs should range from 4 to 52 weeks, depending on the disease being studied. Control limits were chosen accordingly.

4. TYPES OF CHANGES IN THE PROCESSES

In this study, we examined four types of changes in the disease series. The four types of changes are illustrated below:

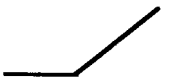
(i) The step shift in the level of the series

(ii) The impulse shift



(iii) The spike shift

(iv) The trend shift



Step, impulse and spike shifts of height one, two and three times σ_a, the standard deviation of the residuals of the ARIMA model, were studied. Impulse shifts of 4, 8, 16 and 24 weeks duration and trend shifts ranging from 0·1 σ_a to 1·0 σ_a per week were studied.

5. SIMULATION OF AVERAGE RUN LENGTHS

5.1. The Recursive Relationship of the Residuals

The ARIMA model for a disease series contains terms relating the current week to past weeks. For example the ARIMA model for hepatitis A is

$$x_t = x_{t-1} + x_{t-52} - x_{t-53} + a_t - 0{\cdot}86a_{t-1} - 0{\cdot}79a_{t-52} + 0{\cdot}6794a_{t-53}$$

where the x's are the square roots of the number of cases of hepatitis A and the a's are random shocks, normally distributed random variables with mean zero and variance $\sigma_a^2 (a \sim N(0, \sigma_a^2))$. The square root transformation was used in the modelling process so that the normality of the random shocks would be justified. This model contains terms relating the current week, x_t, to the .previous week (x_{t-1} and a_{t-1}), as well as seasonal terms relating to the previous year (x_{t-52}, x_{t-53}, a_{t-52} and a_{t-53}).

Following the approach of Wardell *et al.*, we can derive a recursive relationship for the residuals from one-step ahead forecasts based on the ARIMA models of Williamson and Weatherby. For example, the relationship for hepatitis A is

$$\begin{aligned} R_t = {} & s_t - s_{t-1} - s_{t-52} + s_{t-53} + a_t - 0{\cdot}86a_{t-1} - 0{\cdot}79a_{t-52} + 0{\cdot}6794a_{t-53} \\ & + 0{\cdot}86R_{t-1} + 0{\cdot}79\ R_{t-52} - 0{\cdot}6794R_{t-53} \end{aligned}$$

where the s's are the levels of the shift. This equation relates the value of the residual at the present time period, R_t, to present values of the shift and the random shock, s_t and a_t, and to past values of the shift (s_{t-1}, s_{t-52} and s_{t-53}), the random shocks (a_{t-1}, a_{t-52} and a_{t-53}) and the residuals (R_{t-1}, R_{t-52} and R_{t-53}). The presence of seasonal terms in the ARIMA model for hepatitis A leads to the presence of seasonal terms in this relationship.

5.2. The Dynamic Behaviour of the Residuals

The above expression for the residuals illustrates the dynamic response of the residuals to a shift in the process level described by Wardell *et al.* ARIMA models are adaptive forecasting models.

When an ARIMA process undergoes a shift, the expected value of the forecast converges to a new equilibrium level. The expected values of the residuals from the one-step ahead forecasts also converge to a new equilibrium level, smaller in magnitude than the level of the shift. In the above expression, the effect of the shift quickly disappears due to the $s_t - s_{t-1}$ term. The expected values of the residuals converge because the absolute values of the coefficients of the random shock and residual terms are less than one.

The seasonal terms in our ARIMA model introduce an echo effect in the behaviour of the residuals. In the recursive relationship of the residuals, the shift value from a season in the past, s_{t-52}, appears, but it is reduced by the previous value of the shift, s_{t-53}. Values of the random shocks and residuals from a season in the past also appear, but with coefficients that are less than one in absolute value. After experiencing this echo shift, the expected values of the residuals converge to a value that is smaller in magnitude than the shift.

5.3. Simulation

The normally distributed a_t's were simulated. We then made use of the recursive relationship of the residuals to generate R_t, the residual at time t, and subsequently the control statistic based on the residual. The run length, the number of steps taken until the control statistic exceeded the control limits, was noted. The procedure was repeated 10,000 times, and the mean of those 10,000 run lengths was reported as the average run length.

6. ANALYTICAL CALCULATION OF AVERAGE RUN LENGTHS

6.1. The Distribution of the Residuals

In addition to the simulation method described in the previous section, we developed an analytical method for calculating the ARLs of our control charts. Recognizing the recursive relationship for the residuals as a difference equation for the unknown residual R_t, we found a series solution for R_t. The series solution is $R_t = a_t + s_t + \sum_{j=1}^{t} \psi_j s_{t-j}$, where the ψ coefficients depend on the ARIMA model. Since $a_t \sim N(0, \sigma_a^2)$, we have $R_t \sim \mathrm{N}(\mu_t, \sigma_a^2)$, where $\mu_t = s_t + \sum_{j=1}^{t} \psi_j s_{t-j}$. We used this distribution of R_t to calculate ARLs for each of the four control charts as discussed in the next four sections.

6.2. ARLs for the Shewhart Control Chart

Since the run length (RL) of a control chart is a non-negative discrete random variable, we can write $\mathrm{ARL} = \sum_{t=1}^{\infty} P(\mathrm{RL} > t)$ where $P(\mathrm{RL} > t)$ is the probability that the run length exceeds t. For the Shewhart chart based on the one-step ahead residuals, R_t, from the ARIMA model, with control limits at $\pm$ CL, we have

$$P(\mathrm{RL} > t) = P(|R_0| \leqslant \mathrm{CL}) P(|R_1| \leqslant \mathrm{CL}) \ldots P(|R_{t-1}| \leqslant \mathrm{CL}) = \prod_{i=0}^{t-1} P(|R_i| \leqslant \mathrm{CL}).$$

Thus

$$\mathrm{ARL} = \sum_{t=1}^{\infty} \prod_{i=0}^{t-1} P(|R_i| \leqslant \mathrm{CL}).$$

Using the fact that $R_i \sim N(\mu_i, \sigma_a^2)$, where $\mu_i = s_i + \sum_{j=1}^{i} \psi_j s_{i-j}$, we can calculate the $P(|R_i| \leqslant \mathrm{CL})$ and subsequently the ARL for the Shewhart chart for each disease.

The infinite sum in the above expression for the ARL of the Shewhart chart must be approximated by a finite sum in the actual calculation. In the program written for this calculation, the sum was terminated when $\prod_{i=0}^{t-1} P(|R_i| \leqslant \mathrm{CL}) < 10^{-5}$.

6.3. ARLs for the Moving Average Control Chart

The ARL of the MA chart with span two is calculated in a manner similar to that of the Shewhart chart. The probabilities involved are more complicated due to the nature of the MA chart statistic, but the basic approach is the same.

6.4. ARLs for the Exponentially Weighted Moving Average Control Chart

For the EWMA control chart based on the one-step ahead residuals from the ARIMA model, a Markov chain approach similar to that of Lin and Adams is used. The distribution of the residuals derived in Section 6.1 is used to calculate the transition probabilities in the Markov chain. The approach is complicated in our problem by the seasonality of our models, but a modification of the Lin and Adams technique can be used to calculate the ARLs for the EWMA chart.

6.5. ARLs for the Cumulative Sum Control Chart

The average run lengths for the CUSUM control chart were calculated using an approach similar to that used for the EWMA chart. The only difference in the approaches is the use of one CUSUM chart to detect upward shifts in the series and another CUSUM chart to detect downward shifts in the series. The overall average run length for the CUSUM chart was calculated in the usual way, using

$$\frac{1}{\mathrm{ARL}} = \frac{1}{\mathrm{ARL}_{\mathrm{up}}} + \frac{1}{\mathrm{ARL}_{\mathrm{down}}}.$$

The ARLs calculated using the above analytical methods agree quite well with those from the simulation described earlier.

7. RESULTS

The general effect of correlation of the data series on the ARLs of control charts has been described previously by Johnson and Bagshaw,[9] Bagshaw and Johnson,[10] Montgomery and Mastrangelo,[11] Wardell *et al.*[12] and VanBrackle and Reynolds.[13] In addition, the properties of control charts based on residuals from one-step ahead forecasts of some simple time series models have been described by Superville and Adams,[14] Runger *et al.* [15] and Wardell *et al.*[16] In general the effect of positive correlation of the data leads to shortened in-control ARLs (a higher false alarm rate) of control charts based on the data, if the usual control limits are used. If the control limits are adjusted to take the correlation into account and achieve the desired in-control ARL, the effect of the correlation is to delay detection of a shift in the level of the process.

Control charts based on residuals of one-step ahead forecasts of time series models exhibit somewhat different behaviour. The independence of the residuals yields in-control ARLs which are identical to those for control charts based on independent data. However, the adaptive behaviour of the residuals discussed above leads to delayed detection of shifts in the level of the

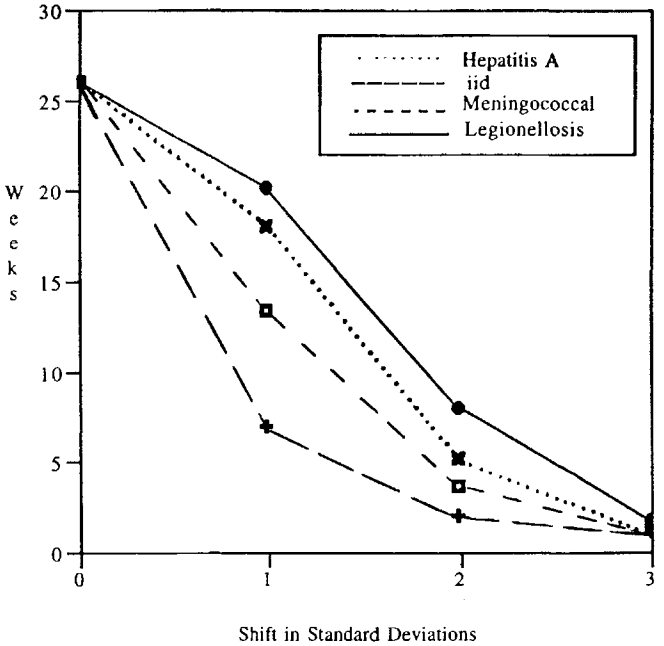


Figure 1. ARLs for the Shewhart chart for a step shift

process. Control charts based on residuals have a window of opportunity for detecting process shifts. The convergence of the expected value of the residuals to a value lower than the shift value gives charts based on residuals a decreased probability of detecting process shifts after the first few time periods following the shift. The seasonality of our ARIMA models does give the chart based on residuals a second or third chance to detect a process shift one or two seasons after the shift, but these second and third chances come too late to be of help in the rapid detection of shifts. Consequently, control charts based on residuals from ARIMA models can have out-of-control ARLs that are unacceptably high.

The results shown in the following figures illustrate the influence of the adaptive behaviour of the residuals on the ARLs of control charts based on the one-step ahead forecast residuals.

Figure 1 shows the ARLs calculated for the Shewhart chart for the independent, identically distributed (i.i.d) data for which the chart was designed and for the residuals from the models for hepatitis A, legionellosis and meningococcal infections. The control limits are chosen to give an in-control ARL of 26 for all of the models. The shift is the step shift measured in units of σ_a, the standard deviation of the residuals from the ARIMA model. The influence of the adaptive behaviour of the residuals is clear; the ARLs for each of the disease models is higher than that for the i.i.d. series at each of the shift levels. For a shift of one standard deviation, the ARLs for the disease models are roughly two to three times the ARL of the i.i.d. series. For a shift of two standard deviations, the disease model ARLs range from two to five times the i.i.d. ARL. This effect is much smaller for the shift of three standard deviations, since the likelihood of detecting such a large shift within the first few time periods after the shift is so high that the adaptive behaviour of the residuals can have little effect.

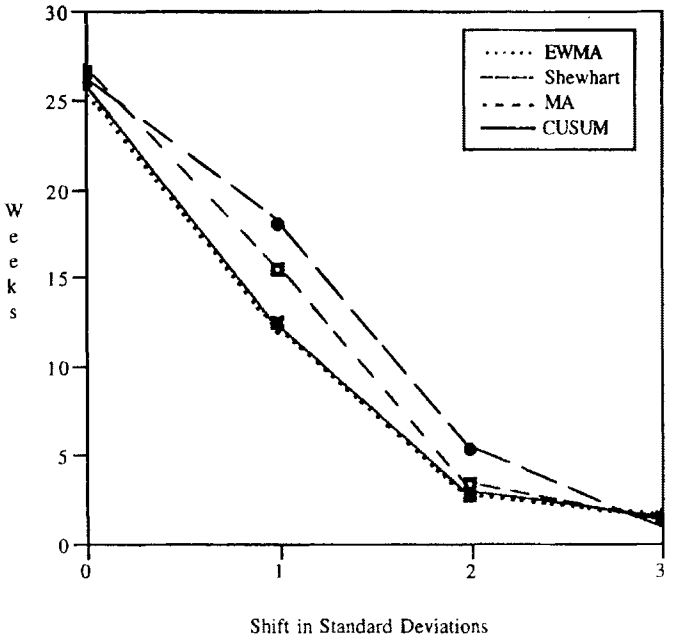


Figure 2. ARLs for hepatitis A for a step shift

It is interesting to note the differences in the behaviours of the different models. The model for legionellosis contains no seasonal terms. As a result, the adaptive behaviour of the residuals for this model is strongest, with no echo effect due to seasonality. Consequently, the Shewhart chart based on the residuals from the legionellosis model has the worst ARL behaviour of the models shown in Figure 1.

As discussed in Section 5.2, the form of the model for hepatitis A causes the residuals to adapt rather rapidly. The seasonal behaviour gives the residuals a second chance to detect the shift before the residuals again adapt. Thus, the Shewhart chart based on the residuals from the hepatitis A model has slightly lower ARLs than the chart for the legionellosis model.

The meningococcal infections model has weaker adaptive behaviour than the hepatitis A model. The residuals from this model do not adapt very rapidly until after the first season (52 weeks). Since the residuals from this model have the weakest adaptive behaviour, the Shewhart chart based on the residuals from the meningococcal infections model has the best ARL characteristics of the disease models in Figure 1.

The remaining results in Figures 2 to 4 are shown for the hepatitis A model. Results are shown for all four types of shift in the level of the process and for all four types of control charts. The other disease series have similar results, differing only in severity according to the complexity and the degree of seasonality of the model.

Figure 2 shows the ARLs for the step shift. Note that the ARLs are all greater than those for the i.i.d. series in the previous figure. It is clear from this figure that the control charts which are designed to accumulate evidence of a shift in the process, the EWMA and the CUSUM charts, have superior detection ability in this situation. The EWMA and CUSUM ARL characteristics

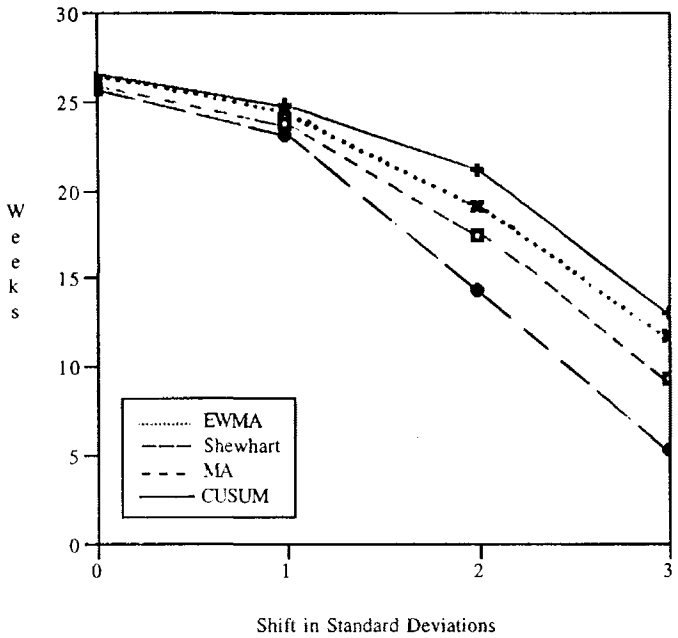


Figure 3. ARLs for hepatitis A for a spike shift

are almost identical, and their lines on the figure are nearly indistinguishable. The MA chart, with its short term accumulation of evidence, outperforms the Shewhart chart. The Shewhart chart has the longest detection times of all, since the adaptive behaviour of the residuals quickly disguises the shift in the process level.

The relative ARL characteristics of the control charts for an impulse shift are similar to those of Figure 1. The ARLs are slightly lower than those for the step shift. The impulse shift consists of two shifts in the process level, a shift up followed by a shift back down to the original level some weeks later. This double shift gives the control charts two opportunities to detect the shift in the process level. The residuals must adapt themselves to both shifts, and the control charts detect the change slightly more quickly than for the step shift.

Figure 3 shows the ARL behaviour for the four control charts for a spike shift, a shift of one week duration. Note the reversal of the ordering of the control charts in this figure. While all of the charts perform rather badly for this type of shift, the accumulating charts (MA, EWMA and CUSUM) perform worse than the Shewhart chart. The spike shift gives no opportunity for the accumulation of evidence of a shift in the process level. As the size of the shift increases, the difference in the detection ability of the charts becomes even more pronounced. The region of large shift is where the Shewhart chart typically performs best. With no evidence to accumulate, the other charts cannot compete with the Shewhart chart.

Finally, Figure 4 illustrates the ARL characteristics of the control charts for a trend shift. The shift level is measured from 0·1 σ_a per week to 1 σ_a per week. For the smaller values of the trend shift, less than 0·5 σ_a per week, the ordering of the charts is as it was for the step shift. The EWMA and CUSUM charts are again nearly identical in their ARL characteristics. The difference

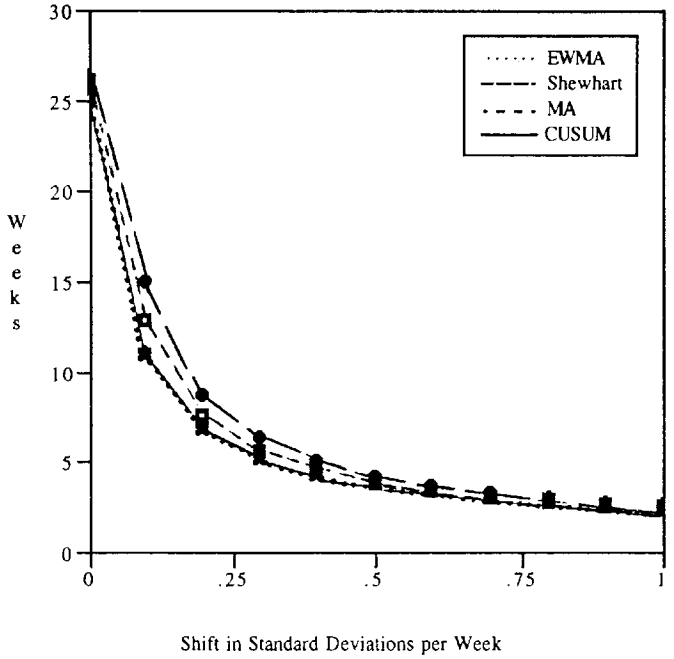


Figure 4. ARLs for hepatitis A for a trend shift

between the ARLs of all the charts is small, and they all perform well. For larger values of the trend shift, the Shewhart chart performs slightly better than the others, but all of the charts detect such large trend shifts quite well.

8. DISCUSSION

The ARL behaviour described above applies to the seven successfully modelled national level disease series only. Much work remains to be done, both for those diseases already modelled and for the other diseases monitored in the NNDSS. For the diseases already modelled, the effects of varying the control chart parameters remain to be studied. The combined Shewhart–EWMA chart was shown by Lin and Adams to be more effective than the traditional control charts for simpler non-seasonal models. The usefulness of this chart for our more complicated seasonal models remains to be evaluated. In addition, the characteristics of control charts applied to state level data and to those disease series for which no good model could be found need to be examined.

Other approaches to the modelling and detection problem also remain to be tried. A combined spatial and time series modelling approach may yield useful results by detecting not only the time of the aberration in a disease series, but also the location of that aberration. Bayesian modelling, change point theory and neural net modelling approaches have yet to be tried on this problem, but may have much to offer in this context.

REFERENCES

1. Williamson, G. D. and Weatherby Hudson, G. 'A monitoring system for detecting aberrations in public health surveillance reports', *Statistics in Medicine*, **18**, 3283–3298 (1999).
2. Box, G. E. P. and Jenkins, G. M. *Time Series Analysis*: *Forecasting and Control*, Holden-Day, San Francisco, 1976.
3. Alwan, L. C. and Roberts, H. V. 'Time-series modeling for statistical process control', *Journal of Business and Economic Statistics*, **6**, 87–95 (1988).
4. Wardell, D. G., Moskowitz, H. and Plante, R. D. 'Run-length distributions of special-cause charts for correlated processes', *Technometrics*, **36**, 3–16 (1994).
5. Lin, W. S. W. and Adams, B. M. 'Combined control charts for forecast-based monitoring schemes', *Journal of Quality Technology*, **28**, 289–301 (1996).
6. Roberts, S. W. 'Control chart tests based on geometric moving averages', *Technometrics*, **1**, 239–250 (1959).
7. Hunter, J. S. 'The exponentially weighted moving average', *Journal of Quality Technology*, **18**, 203–210 (1986).
8. Page, E. S. 'Continuous inspection schemes', *Biometrika*, **41**, 100–114 (1954).
9. Johnson, R. A. and Bagshaw, M. 'The effect of serial correlation on the performance of CUSUM tests', *Technometrics*, **16**, 103–112 (1975).
10. Bagshaw, M. and Johnson, R. A. 'The effect of serial correlation on the performance of CUSUM tests II', *Technometrics*, **17**, 73–80 (1976).
11. Montgomery, D. C. and Mastrangelo, C. M. 'Some statistical process control methods for autocorrelated data', *Journal of Quality Technology*, **23**, 179–193 (1991).
12. Wardell, D. G., Moskowitz, H. and Plante, R. D. 'Control charts in the presence of data correlation', *Management Science*, **38**, 1084–1105 (1992).
13. VanBrackle, L. N. and Reynolds, M. R., Jr. 'EWMA and CUSUM control charts in the presence of correlation', *Communications in Statistics – Simulation and Computation*, **26**, 979–1008 (1997).
14. Superville, C. R. and Adams, B. M. 'An evaluation of forecast-based quality control schemes', *Communications in Statistics – Simulation and Computation*, **23**, 645–661 (1994).
15. Runger, G. C., Willemain, T. R. and Prabhu, S. 'Average run lengths for CUSUM control charts applied to residuals', *Communications in Statistics – Theory and Methods*, **24**, 273–282 (1995).
16. Wardell, D. G., Moskowitz, H. and Plante, R. D. 'Run-length distributions of residuals control charts for autocorrelated processes', *Journal of Quality Technology*, **26**, 308–317 (1994).

STATISTICS IN MEDICINE
Statist. Med. **18**, 3321–3336 (1999)

ESTIMATING GENETIC INFLUENCE ON DISEASE FROM POPULATION-BASED CASE-CONTROL DATA: APPLICATION TO CANCERS OF THE BREAST AND O ARY

GAIL GONG * AND ALICE S. WHITTEMORE

Stanford University School of Medicine, Department of Health Research and Policy, Stanford, CA 94305-5405, U.S.A.

SUMMARY

We describe genetic mixture models and goodness-of-fit statistics for evaluating the joint effects of genetic and environmental factors on the risk of chronic diseases. We focus particularly on situations wherein the gene(s) of interest play roles in several diseases, and death due to one disease can censor the occurrence of others. We use the methods to investigate the risks of cancers of the breast and ovary associated with germline mutations of BRCA1, using data pooled from three population-based U.S. case-control studies of ovarian cancer. We evaluate the goodness-of-fit of the genetic models by comparing the predicted numbers of diseased mother–daughter and sister–sister pairs to the numbers observed. We also use simulations to examine the performance of estimates obtained from such complex mixture models, and the contribution of control families to the precision of parameter estimates. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

A gene is a sequence of base pairs on a chromosome that contains information needed for cellular activities, such as protein synthesis. Throughout human history, mutations have occurred in the genes of the germ cells (that is, the oocytes or spermatocytes) involved in reproduction. Such mutations, which exist in all cells of the offspring arising from the germ cell, are called *germline* mutations. New molecular genetic techniques have helped to identify several genes with germline mutations that increase susceptibility to specific cancers in the individuals who carry them. These discoveries have generated considerable controversy about the advisability of clinical testing for the mutations. The controversy arises from the limited preventive strategies available to mutation carriers, and the possibility that they and their offspring might be denied health insurance, life insurance, or employment because of their cancer risks. In light of this controversy and the intensity of public interest in the issue, it is important that clinicians, public health workers and patients have information on the prevalence of specific germline mutations in the general population, the age-specific cancer risks associated with having such mutations, and the proportion of site-specific cancers attributable to the mutations.

* Correspondence to: Gail Gong, Stanford University School of Medicine, Department of Health Research and Policy, Stanford, CA 94305-5405, U.S.A. E-mail: gailgong@calweb.com

CCC 0277–6715/99/233321–16$17.50

These questions are particularly relevant to cancers of the breast and ovary associated with germline mutations of the gene BRCA1, located on chromosome 17. A woman who inherits from either parent a mutated copy (or *allele*) of this gene has elevated risk for both cancers. The proportion of women who carry a mutation is thought to be quite small, perhaps one in 500 or one in 1000 people. The gene was identified by examining small sections of DNA whose chromosomal locations are known (called *markers*) in the rare families containing multiple cases of breast cancer and ovarian cancer. Within such high-risk families, women affected with breast cancer or ovarian cancer were found to share the same alleles of markers near BRCA1, while unaffected women carried different alleles.[1]

The cancer experience of women in these high-risk families has been used to evaluate the lifetime risks of breast cancer and ovarian cancer associated with BRCA1 mutations.[2] However, these families were ascertained because of their excessive breast and/or ovarian cancer occurrence, and therefore the cancer risks of their members may not represent those of BRCA1 mutation carriers in the general population. Although the investigators attempted to adjust statistically for this potential selection bias, there remains the possibility that risks are overestimated.

In the quest for obtaining unbiased estimates, testing the general population for mutations of BRCA1 is not feasible for the following reasons. First, since mutations are rare, thousands of people would need to be tested, and testing is costly. Also, testing is insensitive; current testing methods fail to detect as many as 30 per cent of all carriers. Because of these problems, estimates of the proportion of the population with BRCA1 mutations and the lifetime cancer risks associated with having a mutation have been obtained not by observing mutations of BRCA1 in individuals but by statistically modelling the patterns of cancer occurrence in population-based samples of families.[3, 4]

In this paper, we describe the statistical methods used to obtain such estimates, and for illustration, we apply these methods to pooled data from three population-based U.S. case-control studies of ovarian cancer. The data are the reported times to breast and ovarian cancer or to censoring of mothers and sisters of ovarian cancer cases and controls. BRCA1 mutations have been estimated to account for 88 per cent (95 per cent CI: 74–97 per cent) of families containing multiple cases of both breast and ovarian cancer and no male breast cancer.[5] Because of the ovarian cancer case-control design of the three studies, and because breast cancer is more common than ovarian cancer, most of the high-risk families (that is, those with two or more cases of breast or ovarian cancer) contain at least one woman with each of the two malignancies and are thus highly likely to carry mutations of BRCA1, rather than some other gene. Accordingly, we shall ignore the possible presence in these families of other genes predisposing to breast or ovarian cancer (see Section 5 for further comment on this point).

Three features of the data complicate the analysis. First, as was mentioned above, the actual mutation status (that is, the *genotype*) of each family member is unknown, and therefore it is necessary to fit a mixture model summing over the unobserved genotypes of family members. Second, BRCA1 strongly influences risks of both breast and ovarian cancer. A woman having either disease has increased risk of dying from that disease and thus censoring her time to the other disease. We call this effect *early censoring*, and under it, the usual assumption of independence among survival and censoring times is violated. Alternative arguments are required to justify the likelihood of the family's censored data. Third, families are not randomly sampled, but instead recruited because a family member residing in a defined geographical area during a certain period of time either developed ovarian cancer (the case families) or was free from ovarian cancer (the control families). Such family ascertainment requires replacing the likelihood with a conditional likelihood.

Theoretical underpinnings aside, how do our methods perform in practice? Mixture models notoriously require huge amounts of data to get good estimates. Given the numbers of families we might expect to collect in practice, can we unravel the effects of genotype on diseases by observing just disease patterns in families? Also, what is the effect of early censoring on performance, again given these practical numbers of families? In the presence of early censoring, what is the performance if we unwittingly focus exclusively on one disease when the mutations of interest cause more than one potentially fatal disease? In our study of ovarian cancer, using a one-disease model on just ovarian cancer is suspect because we would not be taking into account possible early censoring from breast cancer. If we used this incorrect model, what would be its performance? All these questions we examine using simulations.

Section 2 describes the basic genetic model, gives its implications for the joint distribution of times to the diseases of interest, introduces the complications arising from censoring of times to disease, and describes the assumptions needed to deal with such censoring and their biological interpretations. Section 3 illustrates the methods by applying them to the familial occurrence of cancers of the breast and ovary in pooled data from three case-control studies of ovarian cancer. Section 4 presents results from simulations directed toward the questions raised in the previous paragraph, and Section 5 concludes with a brief discussion of the methods' strengths and weaknesses.

2. THE BASIC MODEL

For concreteness, we consider a gene with major effects on two diseases, such as BRCA1 which affects both breast cancer and ovarian cancer. Restriction to one disease and extension to more than two diseases will be apparent. Suppose we have a family with N members, and for the nth member we observe a covariate vector z_n and her time t_{nj} to disease j, $j = 1, 2$. We want to write the probability density function of $\mathbf{t} = (t_{11}, t_{12}, \ldots, t_{N1}, t_{N2})$ conditional on covariates $\mathbf{z} = (z_1, \ldots, z_N)$, allowing for the possibility that disease rates depend on genotypes as well as on covariates. The genotype g_n of the nth family member is a function of her alleles at the gene involved in the disease. For convenience, we introduce the vector $\mathbf{g} = (g_1, \ldots, g_N)$ of genotypes for the family members, which we call the *family genotype*. We assume (i) times to disease among family members are independent, given their genotypes and covariates, and (ii) an individual's times to the two diseases are independent given her genotype and covariates. Since we do not observe genotypes, the probability density function of $\mathbf{t}$ conditional on covariates $\mathbf{z}$ is the following mixture over the distribution $p(\mathbf{g})$ of family genotypes $\mathbf{g} = (g_1, \ldots, g_N)$:

$$\sum_{\mathbf{g}} p(\mathbf{g}) \prod_{j=1}^{2} \prod_{n=1}^{N} f_{jg_n}(t_{nj}|z_n). \tag{1}$$

Here $f_{jg}(t_j|z)$ is the probability density of time t_j to disease j given genotype g and covariates z. To complete model (1) we need to specify the distribution p of the family genotype together with the possible values an individual's genotype g_n can take, and we need to specify the penetrance functions f_{jg} that determine the joint effects of genotypes and covariates on times to disease.

To illustrate the specification of p, consider a dominant Mendelian mode of inheritance, with g an indicator equal to one if an individual inherits at least one mutated allele of a single predisposing gene, and zero otherwise. By contrast, if the mutation acts on disease risks in a recessive way, $g = 1$ only if the individual inherits two mutated alleles, one from each parent. In both instances,

the family genotype $\mathbf{g} = (g_1, \ldots, g_N)$ takes values in the lattice in N-dimensional space consisting of all 2^N points with co-ordinates equal to zero or one, and $p(\mathbf{g})$ depends on the relationships of the family members, the unknown relative frequency ρ of alleles that are mutated, and the laws of inheritance.[6] For definiteness, we shall restrict attention to such dichotomous genotypes, with $g = 1$ indicating the high-risk genotype.

We assume specific parametric forms for the penetrance densities $f_{jg}(t|z)$ for the two diseases, $j = 1, 2$, given genotype g, $g = 0, 1$, and covariates z. In some applications it is useful to assume the hazard functions $\lambda_{jg}(t|z)$ of these densities to be piecewise constant within each of the same A age intervals and within the same C categories of the covariates. The value of λ_{jg} in the ath age interval for an individual in covariate category c is modelled by

$$\lambda_{jgac} = e^{\eta_{jgac}}, \quad j = 1, 2; \ g = 0, 1; \ a = 1, \ldots, A; \ c = 1, \ldots, C$$

where η_{jgac} specifies the joint effects of genes, age and environment on disease j. When the η are independent of the genotypes g, (1) reduces to a Poisson regression model[7, 8] for disease versus covariates. We will call this Poisson model the non-genetic model. In other applications, smooth parametric forms for the $f_{jg}(t|z)$ may be more appropriate. In either case, the unknown parameters in the model are the mutation frequency ρ and the parameters governing the densities $f_{jg}(t|z)$.

We wish to accommodate censored data. That is, instead of observing times t_j to disease, we observe only (u_j, ε_j), where u_j is the minimum of time t_j to disease j and time to censoring, and ε_j equals 1 if $u_j = t_j$ and 0 otherwise. If we assumed that (iii′) conditional on covariates, an individual's time to censoring is independent of her times to the two diseases and independent of her genotype, we could use the usual procedure of replacing the density f_{jg} with the survival function S_{jg} when the time to disease is censored to obtain the joint density of the family's censored data, up to a proportionality constant

$$\sum_{\mathbf{g}} p(\mathbf{g}) \prod_{j=1}^{2} \prod_{n=1}^{N} f_{jg_n}(u_{nj}|z_n)^{\varepsilon_{nj}} S_{jg_n}(u_{nj}|z_n)^{1-\varepsilon_{nj}}. \tag{2}$$

However, we want to allow for less restrictive assumptions on censoring. In particular, we wish to allow death due to one of the two diseases to censor the time to the other disease. Thus instead we assume that family members' progression through the disease and censoring states forms a continuous time Markov chain. The situation is shown graphically in Figure 1, where for simplicity covariates are omitted. All individuals are assumed to be disease-free at birth. Thereafter they are at risk of developing disease j at genotype-specific hazard rates $\lambda_{jg}(t)$, $j = 1, 2$, $g = 0, 1$. A disease-free person may also be censored by death or study termination at rate $\mu_0(t)$. Given the occurrence of either disease 1 or 2, a person is at risk of developing the other disease at the disease-free rates $\lambda_{2g}(t)$ and $\lambda_{1g}(t)$, or of being censored at rates $\mu_1(t)$ and $\mu_2(t)$, respectively. Individuals with both diseases are censored at rate $\mu_3(t)$. We assume (iii) given her covariates and her current state, an individual's time to censoring is independent of her times to the remaining diseases and independent of her genotype. (In particular, her genotype is assumed to be unrelated to risk of fatal diseases that are omitted from the model.) We show in the Appendix that (2) remains proportional to the probability of the family's censored survival data even when assumption (iii′) is replaced by the weaker assumption (iii).

For rare diseases, obtaining results from a random sample of censored survival times from the general population of families, either unconditionally or conditionally on the values of their covariates, is usually inefficient or unfeasible. Instead, sampling is based on some function of

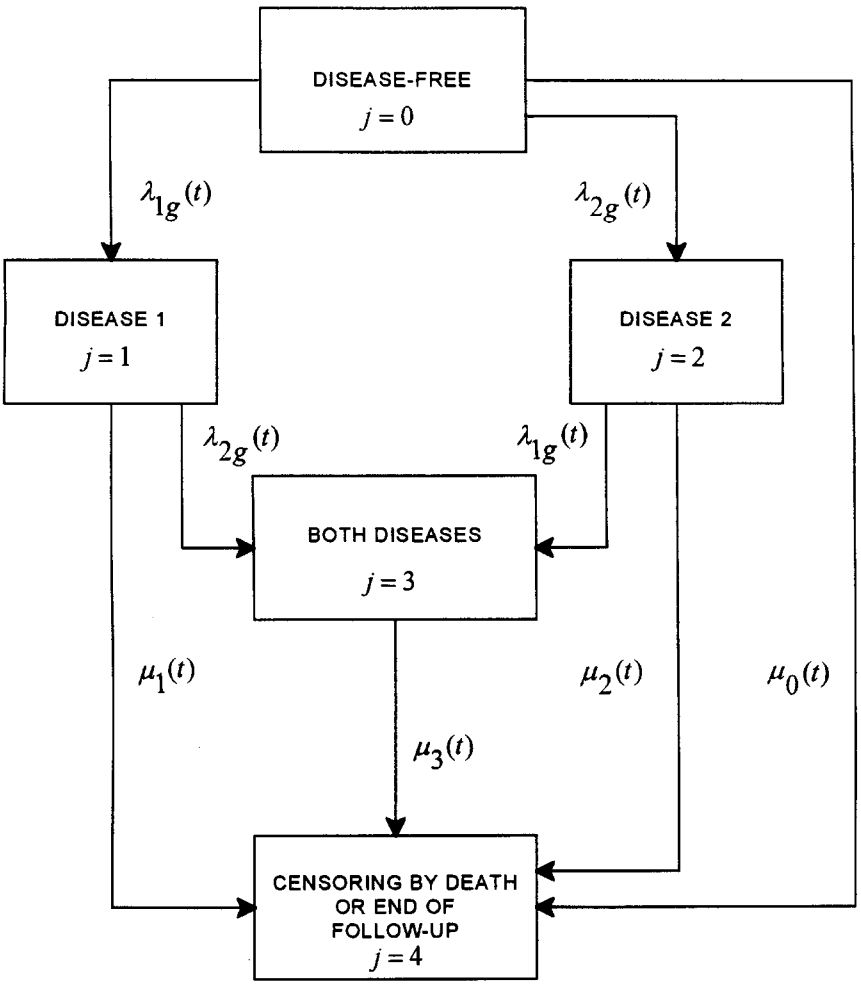


Figure 1. Continuous time Markov chain depicting the occurrence of two diseases in carriers ($g = 1$) and non-carriers ($g = 0$) of genetic mutations. All individuals are assumed to be disease-free at birth. Thereafter an individual is at risk of developing disease j at genotype-specific hazard rates $\lambda_{jg}(t)$. She may also be censored by death or study termination at rate $\mu_0(t)$. Given the occurrence of either disease 1 or 2, a person is at risk of developing the other disease at the disease-free rates $\lambda_{2g}(t)$ and $\lambda_{1g}(t)$, or of being censored at rates $\mu_1(t)$ and $\mu_2(t)$, respectively. Given her genotype, an individual's transition times from one state to another are assumed mutually independent

family survival data, such as the data of one of its members, as in the ovarian cancer case-control example.[9] Tosteson *et al.*[10] show that the distribution of a family's responses under case-control sampling is approximately equal to its distribution conditional on the proband's disease status. To avoid explicit modelling of the censoring distributions, we condition not only on the proband's disease status but also on her age at disease or censoring. When probands (indexed $n = N$) are selected for the presence or absence of disease $j = 2$, the probability of the resulting data is

proportional to the ratio of (2) divided by the factor

$$\sum_{g_N} p(g_N) f_{2g_N}(u_{N2}|z_N)^{\varepsilon_{N2}} S_{2g_N}(u_{N2}|z_N)^{1-\varepsilon_{N2}}. \tag{3}$$

The likelihood of all the case-control family data is the product of the terms (2) divided by (3) taken over all case and control families. To obtain likelihood-based estimates and measures of dispersion for the mutation frequency ρ and the penetrance parameters η_{jgac}, we maximize this likelihood iteratively using Powell's method (see Press *et al.*,[11] pp. 412–420), and by inverting the observed information matrix.

3. EXAMPLE: CANCERS OF THE O ARY AND BREAST

For illustration, we apply the methods to the familial occurrence of cancers of the breast and ovary in pooled data from three case-control studies of ovarian cancer,[12–14] involving family data for 922 ovarian cancer cases and 4951 controls. The individuals who are the cases or controls we call *probands*. Each proband reported the vital statuses of her mother and her sisters, and their ages at death or at the time of interview. She also reported any occurrences of breast or ovarian cancer in these relatives, or of breast cancer in herself. If the proband reported such a cancer, she also reported the age when the cancer was diagnosed. Such information was gathered on both full sisters and half sisters without distinguishing the two. (Coding half sibs as full sibs introduces bias in the parameter estimates which, though likely to be small when there are relatively few half sibs, can be avoided in future studies by gathering the information to distinguish the two.) Information on male breast cancer was not collected. We use 1844 families in our analyses, of which 922 are the case families and 922 are a random sample of the 4951 control families. Including all 4951 control families would have led to huge increases in computing time with little additional precision.[15]

Column 2 of Table I shows the distributions of case and control probands according to their numbers of sisters. Column 3 shows the number of probands who reported a prior diagnosis of breast cancer. Columns 4–7 give the numbers of affected mothers and sisters of probands, by sibship size. In row 4, for example, 98 ovarian cancer cases reported having three sisters. Of these, one ovarian cancer case reported having a prior breast cancer, four reported having a mother with breast cancer, and one reported having a mother with ovarian cancer. In addition, four sisters of the 98 probands were reported to have had a diagnosis of breast cancer, and three sisters were reported to have had a diagnosis of ovarian cancer.

To analyse these data, we consider models in which risks of breast cancer ($j=1$) and ovarian cancer ($j=2$) are both determined by the same autosomal dominant gene. Recall that ρ denotes the relative frequency of mutated alleles, collectively labelled as allele D. Let genotype $g=1$ denote carriers of one or two copies of D and $g=0$ denote those with normal genotype dd. Assuming Hardy–Weinberg equilibrium,[6] the marginal probability that an individual has DD, Dd and dd is, respectively, ρ^2, $2\rho(1-\rho)$ and $(1-\rho)^2$, and so the probabilities of the low-risk ($g=0$) and high-risk ($g=1$) genotypes are, respectively, $p(0)=(1-\rho)^2$ and $p(1)=1-p(0)$. Given the genotype of her parents, a woman receives one allele from each parent with equal probability and independently. For example, if both parents are heterozygous, with genotypes Dd, then she would inherit genotype DD, Dd and dd with probability $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ respectively. The probability distribution $p(g|g_1,g_2)$ of offspring genotype g given parents' genotypes g_1 and g_2 can be obtained

Table I. Distribution of ovarian cancer case and control probands by number of sisters and reported prevalence of breast and ovarian cancer in their first-degree relatives

Number of Sisters	N	Probands	Affected relatives			
			Mothers		Sisters	
		Affected with breast cancer	Breast	Ovary	Breast	Ovary
Cases						
0	259	5	20	6	—	—
1	274	4	16	10	13	0
2	188	5	11	4	8	0
3	98	1	4	1	4	3
4	53	0	0	1	4	1
5	27	0	3	0	1	1
6+	23	0	1	0	1	0
Total	922	15	55	22	31	6
Controls						
0	261	4	18	1	—	—
1	277	8	10	2	4	1
2	183	1	7	1	6	1
3	104	3	2	1	8	0
4	43	0	3	0	0	0
5	30	0	1	0	1	0
6+	24	0	1	0	0	0
Total	922	16	42	5	19	2

using conditional probabilities and Bayes rule: $p(0|0,0)=1,\ p(0|0,1)=p(0|1,0)=(1-\rho)/(2-\rho)$ and $p(0|1,1)=(1-\rho)^2/(2-\rho)^2$. The probability of the entire family's genotype is therefore

$$p(\mathbf{g};\rho)=p(\mathbf{g})=p(g_1)p(g_2)\prod_{n=3}^{N}p(g_n|g_1,g_2) \tag{4}$$

where g_1 is the genotype of the father, g_2 is the genotype of the mother, and $g_3,\ldots,g_N$ are the genotypes of the $N-2$ offspring.

To model the penetrance densities f_{jg} for the two diseases, we assume that the hazard functions λ_{jg} are identically zero before age 15 years, and thereafter are constant within the three age intervals $[15,40)$, $[40,60)$, $[60,100)$. We do not include covariates, so the most general hazard function we consider allows the hazard $\lambda_{jga}=e^{\eta_{jga}}$ of disease j during age interval a of a woman with genotype g to be any positive number. This we call the general hazard model. It involves 13 unknown parameters: the mutation frequency ρ, and the penetrance vector η with $2\times2\times3=12$ components $\eta_{jga},\ j=1,2,\ g=0,1,\ a=1,2,3$.

The non-genetic model assumes $\lambda_{jga}=\lambda_{ja}$ to be independent of genotype g. Under this model, the joint density (2) of the family data simplifies to

$$\prod_{nj}f_j(u_{nj}|\eta)^{\varepsilon_{nj}}S_j(u_{nj}|\eta)^{1-\varepsilon_{nj}}=\exp\sum_{ja}(\Delta_{ja}\eta_{ja}-\tau_{ja}e^{\eta_{ja}})$$

where Δ_{ja} counts the number of family disease j occurrences that fall in the ath age interval and τ_{ja} sums the times to disease j that family members contribute to the ath interval. Since we have $A=3$ intervals with non-zero hazard, this model involves the six unknown parameters in the penetrance vector $\eta=(\eta_{11},\eta_{12},\eta_{13},\eta_{21},\eta_{22},\eta_{23})$.

The third model we consider is the proportional hazard model

$$\eta_{jga}=\alpha_{jg}+\gamma_{ja},\quad j=1,2;\ g=0,1;\ a=1,2,3;\ \alpha_{10}=\alpha_{20}=0.$$

This model involves nine unknown parameters: the mutation frequency ρ and the eight parameters in the pentrance vector $\eta=(\alpha_{11},\alpha_{21},\gamma_{11},\gamma_{12},\gamma_{13},\gamma_{21},\gamma_{22},\gamma_{23})$. The constants $e^{\alpha_{11}}$ and $e^{\alpha_{21}}$ are the rate ratios for ovarian and breast cancer, respectively, in carriers relative to non-carriers. This model specifies that the breast and ovarian cancer rate ratios in carriers versus non-carriers are independent of age. ariations on this model include those in which the carrier hazard is proportional to the non-carrier hazard for one disease only.

In all model fitting, we avoid parameter constraints by using the transformations $\zeta=\log[\rho/(1-\rho)]$ and $\eta_{jga}=\log\lambda_{jga}$, whose ranges are the entire real line. Since the proportional hazard model is nested in the general hazard model, the likelihood ratio statistic can be used to examine its adequacy. The likelihood ratio statistic cannot be used to test the non-genetic model because it does not have an asymptotic chi-squared distribution when comparing this model to any of the genetic mixture models described above.[16] Similarly, we cannot use likelihood ratio theory to evaluate the adequacy of the general hazard model, because it is not nested in a larger model. Later in this section we describe alternative criteria, efficient score statistics, for evaluating models and we show that the non-genetic model fits the data poorly while the general hazard model provides an adequate fit to the data. Further, the likelihood ratio statistic rejects the proportional hazard model in favour of the general hazard model ($p=0{\cdot}05$). We are led therefore to use the general hazard model.

Table II presents parameter estimates and their standard errors based on the general hazard model. The data contain too few breast cancer cases among the oldest women to estimate the carrier hazard λ_{113}, so we assume its value is known to be 16·81, the value obtained in a breast cancer case-control study.[3] These estimates give a mutation frequency of $\hat{\rho}=0{\cdot}0015$ (with 95 per cent non-simultaneous confidence interval $0{\cdot}0002-0{\cdot}011$) leading to an estimate of the prevalence of mutation carriers in the population $1-(1-\hat{\rho})^2=0{\cdot}3$ per cent (0·04–2·2 per cent). When the estimated penetrance densities are cumulated to age 79 years, we obtain breast cancer probabilities of 73·5 per cent (47·3–89·4 per cent) and 6·8 per cent (5·4–8·6 per cent) in carriers and non-carriers, respectively, and corresponding ovarian cancer probabilities of 27·8 per cent (5·2–73·0 per cent) and 1·8 per cent (1·1–15·8 per cent) in carriers and non-carriers, respectively. These population-based risk estimates are lower than those obtained from the selected high-risk families,[2] suggesting that the cancer experience of these families may not represent that of the general population.

We can use these estimates in Bayes rule to estimate the contribution of BRCA1 mutations to the burden of breast cancer or of ovarian cancer among women of various ages. For example, we estimate that among women aged 15–69 years, 4·2 per cent of all breast cancers and 5·3 per cent of all ovarian cancers are due to these mutations. These estimates are consistent with those obtained by mutation testing of unselected samples of cancer cases.[17, 18]

We conclude this section by evaluating goodness-of-fit of the non-genetic and general hazard models using efficient score statistics, which compare the observed numbers of pairs of affected relatives to the numbers predicted by the model. The efficient score is $S=(O-E)/\mathrm{SE}$, where O and E represent the observed and predicted numbers of affected relative pairs, respectively,

Table II. Maximum likelihood estimates (MLEs) and 95 per cent confidence intervals for gene frequency and hazard rates (CIs) in general hazard model for ovarian cancer case-control data

	Gene frequency ρ			
	MLE	CI		
	0·0014	0·0002–0·0109		
	Hazard rates ($\times 10^3$)			
	Carriers		Non-carriers	
Age (years)	MLE	CI	MLE	CI
Breast cancer				
15–39	4·4	1·9–9·9	0·1	0·05–0·2
40–59	44·0	18·3–105·6	1·0	0·7–1·5
60+	16·8*	—	2·4	1·8–3·3
Ovarian cancer				
15–39	1·6	0·5–5·9	0·02	0·01–0·1
40–59	5·8	1·4–25·2	0·2	0·1–0·4
60+	8·4	0·9–79·1	0·6	0·4–1·1

* Assumed known from Claus *et al.*[3]

where a woman was 'affected' if she developed the given cancer (ovary or breast) by a specified age, and SE represents the standard error of O. Under the null hypothesis that the data were generated by the model, and therefore that the model fits the data, the efficient score S has a standard Gaussian distribution.[19] The test accommodates censoring, the case-control design and the dispersion parameter of the maximum likelihood estimates (MLEs).

Table III shows the results of testing whether the non-genetic and general hazard models fit the ovarian cancer case control data. We can see that the non-genetic model fits poorly. It significantly underpredicts the numbers of affected mother–daughter and sister–sister pairs. In contrast, there were no statistically significant differences between observed numbers of affected pairs and those predicted by the general hazard model.

4. PERFORMANCE OF PARAMETER ESTIMATES

How do these methods perform? We use simulations to study the behaviour of the parameter estimates in the mixture models. In particular, if two-disease censored survival family data do in fact follow the model (2) with ρ, λ_{jga} as the true underlying parameters, then assuming the same model to calculate the MLEs $\hat{\rho}$, $\hat{\lambda}_{jga}$, how do the MLEs compare to the true parameters? While the arguments in the Appendix support the consistency and asymptotic efficiency of the estimates, complex mixture models such as the ones considered here may require huge sample sizes to achieve asymptotic performance. How well do the estimates perform with sample sizes that are available in practice to genetic epidemiologists, and what are the effects of early censoring on parameter estimates under both the correct two-disease and incorrect one-disease models?

We simulated case-control data, generating in each replication of the simulations, 1000 case families and 1000 control families. Each family consisted of a proband, her father, mother and two sisters. Thus there were $N = 5$ members in each family. To generate the family data for a proband with disease 2 (a case), we first generated the case's data. We did this by generating

Table III. Observed (O) and observed minus expected ($O-E$) numbers of relative pairs with breast or ovarian cancer before age 50 years

	Cancer types		
	Breast/breast	Breast/ovary	Ovary/ovary
Mother–daughter			
O*	6	19	6
$O-E$† non-genetic	4·9§	6·4	3·4
general hazard	0·8	−2·0	1·0
Sister–sister			
O	5	15	2
$O-E$ non-genetic	3·8§	6·7‡	0·2
general hazard	−0·2	2·7	−1·1
All pairs			
O	11	34	8
$O-E$ non-genetic	8·7§	13·1§	3·6
general hazard	0·6	0·7	0·1

* Including pairs of relatives in which one member is the proband
† E = expected number of pairs according to fitted model, adjusted for censoring and for the case control design. p-values reflect the dispersion in parameter estimates
‡ $p<0{\cdot}05$
§ $p<0{\cdot}001$

her genotype g_N from the distribution $p(g)$, and then generating one of the three sample paths shown in Table I , and continuing this process until we obtained path $0\to 2\to 4$ or path $0\to 2\to 3\to 4$. To generate data for a control proband, we performed this process until we obtained path $0\to 4$ or path $0\to 1\to 4$. Next, we generated the parents' genotypes g_1, g_2 from the distribution $p(g_1,g_2|g_N)=p(g_1,g_2,g_N)/p(g_N)$ and generated the sisters' genotypes g_3, g_4 independently from $p(g|g_1,g_2)$. Finally, we generated independent sample paths for the mothers and sisters, based on the disease rates conditional on their genotypes, and on the assumed censoring rates. To obtain a sample path for any woman with genotype g, we began by generating times t_1, t_2 and t_4 using hazards $\lambda_{1g},\lambda_{2g}$ and μ_0, respectively, and ordering the three times. If $t_4\leqslant \min(t_1,t_2)$, we assigned her the path $0\to 4$. If $t_4>\min(t_1,t_2)=t_1$ say, we generated an early censoring time t_e using hazard μ_e and then redefined $t_4=t_1+t_e$. We then assigned the woman the path $0\to 1\to 4$ if $t_4\leqslant t_2$, or the path $0\to 1\to 3\to 4$ if $t_4>t_2$.

We assumed the censoring rate for disease-free women to be $\mu_0\equiv 1/70$ women per year, and so their mean time to censoring was 70 years. Also, we assumed diseased women are censored at rates $\mu_1=\mu_2=\mu_3=\mu_e\geqslant 1/70$. When $\mu_e=1/70$, censoring times and disease times are independent, while $\mu_e>1/70$ gives early censoring. We examined the performance of estimates varying the rate of early censoring among four values, $\mu_e=1/70$, 1/30, 1/20, 1/10 in both a two-disease model which includes both breast and ovarian cancer and in a one-disease model which includes just ovarian cancer. These rates correspond to five-year disease-survival probabilities of 93, 90, 78 and 61 per cent, respectively. arying four disease rates in two models gives eight trials. In each trial, we generated and analysed six replications of case-control data. While more replications would have been desirable, the generation and analysis of replication requires substantial computer time, and with as few as six replications, we can begin to appreciate the behaviour of the parameter estimates.

Table I . Factorization of probability density of sample path for one individual

	Path		
	1. $0 \to 4$; t_4	2. $0 \to j \to 4$; t_j, t_4	3. $0 \to j \to 3 \to 4$; t_j, t_3, t_4
Probability density	$p_{g00}(0, t_4) q_{g04}(t_4)$	$p_{g00}(0, t_j) q_{g0j}(t_j) p_{gjj}(t_j, t_4) q_{gj4}(t_4)$	$p_{g00}(0, t_j) q_{g0j}(t_j) p_{gjj}(t_j, t_3) q_{gj3}(t_3) p_{g33}(t_3, t_4) q_{g34}(t_4)$
Factor	λ: $\exp\left[-\int_0^{t_4} (\lambda_{g1} + \lambda_{g2})(u)\,\mathrm{d}u\right]$	λ: $\lambda_{gj}(t_j) \exp\left[-\int_0^{t_j} (\lambda_{g1} + \lambda_{g2})(u)\,\mathrm{d}u - \int_{t_j}^{t_4} \lambda_{gj'}(u)\,\mathrm{d}u\right]$	λ: $\lambda_{gj}(t_j)\lambda_{gj'}(t_3) \exp\left[-\int_0^{t_j} (\lambda_{g1} + \lambda_{g2})(u)\,\mathrm{d}u - \int_{t_j}^{t_3} \lambda_{gj'}(u)\,\mathrm{d}u\right]$
	μ: $\mu_0(t_4) \exp\left[-\int_0^{t_4} \mu_0(u)\,\mathrm{d}u\right]$	μ: $\mu_j(t_4) \exp\left[-\int_0^{t_j} \mu_0(u)\,\mathrm{d}u - \int_{t_j}^{t_4} \mu_j(u)\,\mathrm{d}u\right]$	μ: $\mu_3(t_4) \exp\left[-\int_0^{t_j} \mu_0(u)\,\mathrm{d}u - \int_{t_j}^{t_3} \mu_j(u) - \int_{t_3}^{t_4} \mu_3(u)\,\mathrm{d}u\right]$

Table . Performance in simulations of parameter estimates for ovarian cancer when breast cancer is (a) included and (b) excluded from analysis

	Minus the logit of mutation frequency	Estimates* (SE)† Minus the log of ovarian cancer rate					
		Carriers, age (yrs)			Non-carriers, age (yrs)		
		<40	40–60	60+	<40	40–60	60+
True value	4·6	6·0	5·0	5·0	10·0	8·0	7·0
Censoring rate after breast cancer							
(*a*) *Including breast cancer*							
1/70	4·5 (0·1)	6·2 (0·1)	5·1 (0·1)	5·0 (0·1)	10·1 (0·2)	7·9 (0·1)	7·0 (0·03)
1/30	4·7 (0·1)	6·1 (0·1)	4·9 (0·1)	5·3 (0·2)	10·2 (0·1)	8·0 (0·1)	7·0 (0·03)
1/20	4·8 (0·1)	6·2 (0·1)	5·0 (0·1)	5·4 (0·2)	10·3 (0·1)	8·1 (0·1)	6·9 (0·06)
1/10	4·7 (0·1)	6·1 (0·1)	5·2 (0·1)	6·1 (0·2)	10·1 (0·1)	7·9 (0·1)	6·9 (0·04)
(*b*) *Excluding breast cancer*							
1/70	5·4 (0·8)	5·8 (0·4)	4·4 (0·6)	4·4 (0·6)	10·5 (0·9)	7·9 (0·2)	6·9 (0·07)
1/30	5·2 (0·8)	6·0 (0·3)	4·7 (0·5)	4·8 (0·7)	10·2 (0·6)	8·6 (0·9)	6·9 (0·08)
1/20	5·3 (0·8)	6·1 (0·4)	4·7 (0·4)	5·8 (1·6)	9·9 (0·2)	8·1 (0·1)	7·0 (0·12)
1/10	5·7 (0·9)	5·8 (0·5)	4·6 (0·5)	8·2 (4·1)	11·8 (2·2)	8·2 (0·4)	6·9 (0·05)

* Mean of six replications

† SE, standard error based on empirical variance in six replications

The results displayed in Table show several trends. First, when we assume the two-disease model, parameter estimates cluster fairly tightly around their true values. The only egregious spot is the estimate for the carrier hazard of the oldest age group; while unbiased for independent censoring, it is biased downward for early censoring, and the bias increases with increasing rate of early censoring. Second, when we assume the one-disease model, excluding breast cancer, the estimates for the mutation frequency are severely underestimated. Moreover all estimates are highly unstable, especially those for the carrier hazards. For a rare disease, such as ovarian cancer, there are not many occurrences of the disease to get a good estimate for the mutation frequency; small estimates of the mutation frequency are associated with high carrier hazards estimates while large estimates of the mutation frequency are associated with low carrier hazards estimates. The two-disease model does not share this instability because the occurrences of breast cancers tend to pin down the mutation frequency estimate.

5. DISCUSSION

We described mixture models for the joint distribution of survival times to disease in families, and its dependence on genetic and environmental factors. The mixing distribution is the joint distribution of family genotypes, and the component distributions give probabilities of times to diseases in a family, conditional on unobserved family genotypes and observed covariates. Disease times can be censored, and disease hazard rates can vary with age. The models extend previous ones used solely to evaluate genetic effects and others used solely to evaluate non-genetic effects. We focused on the ramifications of multiple or pleitropic effects, wherein the same gene mutation

may increase risk for more than one disease. We saw that even in the presence of early censoring, the usual likelihood of the family censored data remains valid.

We applied the methods to case-control data in which families were chosen to be in the study because the proband developed or was free from ovarian cancer and in which survival times to breast and ovarian cancers were collected for family members.

Using simulations, we saw that except under extreme early censoring, the mixture model can give good estimates for the mutation frequency and the penetrance parameters when we have the numbers of families we might expect in practice. On the other hand, the one-disease model performs poorly, giving unstable results for all degrees of early censoring including independent censoring.

Some biological implications of the methods warrant consideration when applying them to data and interpeting the results. The unavailability of individual genotypes at a specific genetic locus is both a strength and a weakness. It is a strength because 'ballpark' estimates of mutation frequency and the associated disease risks can be obtained and used as rough guidelines for public health planning without the considerable expense of obtaining genotypes for huge numbers of people. It is a weakness because it involves the use of mixture models, which require large numbers of families for reliable estimates and confidence limits. In addition, the interpretation of genotype can be problematic. For example, if more than one gene is involved in the diseases of interest and the relevant mutations all have non-negligible frequencies in the population under study, the classification of genotypes becomes more complex than the 'carrier/non-carrier' dichotomy presented here. In principle, one would need to classify an individual's genotype for each of the relevant genes. The hazard rates for the related diseases would then depend on this expanded classification. However, if mutations of each of the relvant genes are sufficiently rare that no family is likely to carry more than one mutation, genotype could still be coded as a carrier/non-carrier dichotomy, with the understanding that not all carrier families carry mutations of the same gene. The hazard rates for disease in carriers must then be interpreted as averages of gene-specific hazard rates.

APPENDIX

We begin by justifying the likelihood function (2) for randomly sampled families. We wish to verify that the probability of a family's disease data is proportional to expression (2). We shall ignore covariates here, however the arguments can be extended to distributions conditional on covariate values. Assumptions (ii) and (iii) imply that, conditional on her genotype, an individual's data form a sample path of a continuous time Markov chain $X(t)$, where $X(t)$ denotes her state at age t. The five states are: disease-free ($j=0$); disease 1 only ($j=1$); disease 2 only ($j=2$); both diseases ($j=3$); and censored by death or termination of observation ($j=4$). We assume that all individuals are disease-free at birth. We also assume sufficient regularity that the probability distribution of $X(t)$ satisfies the forward Kolmogorov equations (Parzen,[20] p. 291). To describe them, we introduce the 5×5 transition probability matrices $\mathbf{P}_g(s,t)$, $g=0,1$, with entries

$$p_{gjk}(s,t)=\Pr[X(t)=k|X(s)=j,g],\quad j,k=0,1,\ldots,4.$$

For each pair of states j,k, $j\neq k$, we also introduce the non-negative transition intensities, defined by

$$q_{gjk}(t)=\lim_{h\to 0}\frac{1}{h}p_{gjk}(t,t+h),\quad j\neq k$$

and let

$$q_{gjj}(t) = -\sum_{k \neq j} q_{gjk}(t).$$

For our problem, the intensities of transition to the three disease states are $q_{g0j}(t) = \lambda_{jg}(t),\ j = 1, 2$, $q_{g03}(t) = 0,\ q_{g13}(t) = \lambda_{2g}(t),\ q_{g23}(t) = \lambda_{1g}(t)$. The rates of transition to the censored state 4 are $q_{gj4}(t) = \mu_j(t),\ j = 0, 1, 2, 3$, independent of genotype g. Typically $\mu_j(t) \geqslant \mu_0(t)$ for $j = 1, 2, 3$, since the occurrence of one or both diseases may increase the death rate. Finally, $q_{gjk}(t) = 0$, for $k < j$. For $j = 1, 2,\ \lambda_{jg}(t)$ is the disease-j hazard function. The matrix $\mathbf{Q}_g(t) = [q_{gjk}(t)]$ is thus

$$\mathbf{Q}_g(t) = \begin{pmatrix} -(\lambda_{1g} + \lambda_{2g} + \mu_0) & \lambda_{1g} & \lambda_{2g} & 0 & \mu_0 \\ 0 & -(\lambda_{2g} + \mu_1) & 0 & \lambda_{2g} & \mu_1 \\ 0 & 0 & -(\lambda_{1g} + \mu_2) & \lambda_{1g} & \mu_2 \\ 0 & 0 & 0 & -\mu_3 & \mu_3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

where we have omitted the argument t. The forward Kolmogorov equations are[20]

$$\frac{\partial}{\partial t}\mathbf{P}_g(s,t) = \mathbf{P}_g(s,t)\mathbf{Q}_g(t) \quad t \geqslant s; \quad \mathbf{P}_g(s,s) = \mathbf{I}_5. \tag{6}$$

Here $\mathbf{I}_5$ is the identity matrix of dimension 5.

The individual's data comprise one of the following three sample paths of $X(t)$: 1. direct transition from the disease-free state 0 to the censored state 4 at age t_4, denoted $0 \to 4; t_4$; 2. transition from state 0 to state j at time t_j, followed by transition to state 4 at time t_4, $(0 \to j \to 4; t_j, t_4)$, $j = 1, 2$; 3. transition from state 0 to state j at time t_j, $j = 1, 2$, followed by transition to state 3 at time t_3, and finally transition to state 4 at time t_4 $(0 \to j \to 3 \to 4; t_j, t_3, t_4),\ j = 1, 2$.

Table I gives the probability density of a sample path in terms of the entries of matrices $\mathbf{P}_g$ and $\mathbf{Q}_g$. Here $j' = 3 - j,\ j = 1, 2$. Solving the differential equations (6) gives

$$\begin{aligned} p_{g00}(s,t) &= \exp\left[-\int_s^t (\lambda_{1g} + \lambda_{2g} + \mu_0)(u)\,\mathrm{d}u\right] \\ p_{gjj}(s,t) &= \exp\left[-\int_s^t (\lambda_{j'} + \mu_j)(u)\,\mathrm{d}u\right] \quad j = 1, 2 \\ p_{g33}(s,t) &= \exp\left[-\int_s^t \mu_3(u)\,\mathrm{d}u\right]. \end{aligned}$$

We substitute these expressions and the appropriate entries of $\mathbf{Q}_g$ into the densities in Table I , and then factor the resulting densities as products of a term involving λ_{1g} and λ_{2g} (the λ-factor) and a term involving $\mu_0, \ldots, \mu_4$ (the μ-factor). Notice that each λ-factor has the form

$$\prod_{j=1}^{2} \lambda_{jg}(u_j)^{\varepsilon_j} \mathrm{e}^{-\int_0^{u_j} \lambda_{jg}(u)\,\mathrm{d}u} \tag{7}$$

where for $j = 1, 2,\ \varepsilon_j = 1$ if the individual enters state j before censoring and $\varepsilon_j = 0$ otherwise, and $u_j = t_j$ if $\varepsilon_j = 1$, with $u_j = t_4$ if $\varepsilon_j = 0$. Notice also that each μ-factor is the conditional probability density that the individual enters the censored state 4 in the interval $(t_4, t_4 + \mathrm{d}t)$, given her disease history prior to t_4. By assumption (iii), this μ-factor does not depend on the individual's genotype g.

Next we consider a family containing more than one individual. Assumption (i) states that, conditional on their genotypes, the individuals' times to transition from one disease state to another are independent. This assumption implies that, conditional on their genotypes, the λ-factor in the joint probability density of their sample paths is the product of the appropriate λ entries of Table I for each of the individuals. However, the μ-factor is the joint density function of their times to censoring, conditional on their prior histories; it factors only if censoring times for the individuals are independent. Summing over the unobserved genotypes, and using the genotype independence of the censoring density, we see that assumptions (i)–(iii) and (7) imply that the probability density of a family's data is proportional to

$$\sum_{\mathbf{g}} p(\mathbf{g}) \prod_{n=1}^{N} \prod_{j=1}^{2} \lambda_{jg_n}^{\varepsilon_{dn}}(u_{nj}) \mathrm{e}^{-\int_0^{u_{nj}} \lambda_{jg_n}(u)\,\mathrm{d}u} \tag{8}$$

which is just (2). The proportionality factor is the μ-factor for the family, that is, the joint probability density of the family's observed times to arrival in the censored state 4, given the family's disease history.

For case-control sampling with probands sampled according to presence or absence of, say, disease 2, we need to divide (8) by the marginal density of each proband's disease-2 data. Integrating (8) with respect to $(t_{jn}, \varepsilon_{jn})$ $j = 1, 2$, $n = 1, \ldots, N-1$ and $(t_{1N}, \varepsilon_{1N})$ gives this probability as

$$\left\{ \sum_{g_N} p(g_N) \lambda_{2g_N}^{\varepsilon_{N2}}(u_{N2}) \mathrm{e}^{-\int_0^{u_{N2}} \lambda_{2g_N}(u)\,\mathrm{d}u} \right\} \text{constant}$$

where the data-dependent constant involves the censoring and is independent of the unknown mutation frequency and penetrance parameters. The expression in brackets is just (3).

REFERENCES

1. Hall, J., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. and King, M. C. 'Linkage of early-onset familial breast cancer to chromosome 17q21', *Science*, **250**, 1684–1689 (1990).
2. Easton, D. F., Ford, D. and Bishop, D. T. For Breast Cancer Linkage Consortium. 'Breast and ovarian cancer incidence in BRCA1-mutation carriers', *American Journal of Human Genetics*, **56**, 265–271 (1995).
3. Claus, E. B., Risch, N. and Thompson, W. D. 'Genetic analysis of breast cancer in the Cancer and Steroid Hormone Study', *American Journal of Human Genetics*, **48**, 232–242 (1991).
4. Whittemore, A. S., Gong, G. and Itnyre, J. 'Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer: results from three U.S. population-based case-control studies of ovarian cancer', *American Journal of Human Genetics*, **60**, 496–504 (1997).
5. Narod, S. A., Ford, D., Devilee, P. *et al.* 'An evaluation of genetic heterogeneity in 145 breast-ovarian cancer families', Breast Cancer Linkage Consortium, *American Journal of Human Genetics*, **56**, 254–264 (1995).
6. Crow, J. F. and Kimura, M. *An Introduction to Population Genetics Theory*, Harper and Row, New York, 1970.
7. Holford, T. R. 'The analysis of rates and survivorship using log-linear models', *Biometrics*, **36**, 299–305 (1980).
8. 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).
9. Neuhaus, J. M. and Jewell, N. P. 'The effect of retrospective sampling on binary regression models for clustered data', *Biometrics*, **46**, 977–990 (1990).
10. Tosteson, T. D., Rosner, B. and Redline, S. 'Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders', *Biometrics*, **47**, 1257–1265 (1991).

11. Press, W. H., Teukolsky, S. A., etterling, W. T. and Flannery, B. P. *Numerical Recipes in C*, Cambridge University Press, 1992.
12. Cancer and Steroid Hormone CASH Study of the Centers for Disease Control and the National Institute of Child Health and Human Development. 'The reduction in risk of ovarian cancer associated with oral-contraceptive use', *New England Journal of Medicine*, **316**, 650–655 (1987).
13. Casagrande, J. T., Louie, E. W., Pike, M. C., Roy, S., Ross, R. K. and Henderson, B. E. 'Incessant ovulation and ovarian cancer', *Lancet*, **2**, 170–173 (1979).
14. Whittemore, A. S., Wu, M. L., Paffenbarger, R. S., Jr., *et al.* 'Epithelial ovarian cancer and the ability to conceive', *Cancer Research*, **49**, 4047–4052 (1989).
15. Whittemore, A. S. and Halpern, J. 'Multi-stage sampling in genetic epidemiology', *Statistics in Medicine*, **16**, 153–167 (1997).
16. Whittemore, A. S. and Gong, G. 'Segregation analysis of case-control data using generalized estimating equations', *Biometrics*, **50**, 1073–1087 (1994).
17. Langston, A. A., Malone, K. E., Thompson, J. D., Daling, J. R. and Ostrander, E. A. 'BRCA1 mutations in a population-based sample of young women with breast cancer', *New England Journal of Medicine*, **334**, 137–142 (1996).
18. Stratton, J. F., Gayther, S. A., Russell, P. *et al.* 'Contribution of BRCA1 mutations to ovarian cancer', *New England Journal of Medicine*, **336**, 1125–1130 (1997).
19. Whittemore, A. S., Halpern, J. and Gong, G. 'Testing covariance structure in multivariate models: application to family disease data', *Journal of the American Statistical Association*, **93**, 518–525 (1998).
20. Parzen E, *Stochastic Processes*, Holden-Day, San Francisco, 1962.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3337–3344 (1999)

AN APPLICATION OF LIFETIME MODELS IN ESTIMATION OF EXPECTED LENGTH OF STAY OF PATIENTS IN HOSPITAL WITH COMPLEXITY AND AGE ADJUSTMENT

JIANLI LI*

Quality and Clinical Resource Utilization, St. Michael's Hospital, University of Toronto, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada

SUMMARY

Expected length of stay (ELOS) of patients in hospital is an important measure in hospital resource utilization management. Previous work has shown that estimation of ELOS is improved using complexity and age adjustment. These improved estimates have the potential to improve the accuracy of estimates of resource use. Recently other authors have applied the linear regression model to make complexity and age adjustments in the estimation of ELOS. However, these estimates using linear regression estimates are likely flawed on the basis that the assumptions regarding the distribution of data for the linear regression model are unjustifiable. The non-normal distributions of most hospital patient discharge data demand that an alternative method be described to provide accurate estimates of ELOS. The purpose of this paper is to describe an alternative method which uses lifetime models to initially estimate the expected length of stay. The paper then provides an approach to estimate the adjusted expected length of stay (AELOS) using several influencing factors by application of lifetime models. Depending on whether or not the proportional hazards assumption is appropriate for the data, the Cox proportional hazards model or the Kaplan–Meier adjustment is recommended. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

A patient's length of stay (LOS) in hospital is one of most common indicators of the efficiency of the utilization of hospital resources. Accurately estimating the expected length of stay (ELOS) for a specific case could assist with discharge planning, monitoring clinical practice and developing benchmarks.

The resource intensity weight (RIW), which is an estimate of the relative resources used by each type of case, is also an indicator of hospital-resource utilization to allow inter-institutional or inter-jurisdictional comparison. The assignment of RIW to an individual patient case is based on the patient group methodologies. These patient group methodologies are case mix group (CMG) and day procedure groups (DPG) in Canada and diagnosis related groups (DRG) in the United States. The RIW is the ratio of the cost of a case in a CMG to the average cost of all cases in the database.

* Correspondence to: Jianli Li, Quality & Clinical Resource Utilization, St. Michael's Hospital, University of Toranto, 30 Bond Street, Toronto, Ontario M5B 1 W8, Canada. E-mail: LIJ@SMH.TORONTO.ON.CA

CCC 0277–6715/99/233337–08$17.50

This kind of patient grouping methodology generates a manageable number of groups and demonstrates statistical homogeneity with respect to either length of stay (LOS) or total resources use.[2] Patients who fall into the same CMG should be expected to have similar LOS or to consume similar volume of resources.

It has been shown that the RIW is closely related to the length of stay.[1–3] Therefore, the expected length of stay (ELOS) is an important component in the estimation of RIW.

Originally, CMG assignments were designed to group acute inpatients with similar clinical and resource utilization characteristics. The patient's most responsible diagnosis was used to assign a case to one of 25 major clinical categories (MCC). Then the case was directed to a medical or surgical partition based on the presence or absence of an operative procedure. Thus, most of the patient cases could be assigned to a specific CMG, based on the MCC and the presence or absence of an operative procedure. Some cases could be further assigned a separate CMG according to the presence or absence of a complication/co-morbidity (CC) and whether the patient's age was greater than a specific age, for example, 70 years old, or not.[2,3]

It has been found that CMGs alone are insufficient to estimate ELOS. Although it has been found that factors such as medical complexity, clinical severity and age may improve the ability to estimate ELOS,[4–8] they are still insufficient to explain all the variances seen in LOS and hospital resource utilization. In spite of this, complexity and age adjustment still remains an essential component of major predictive models. In this view, the Canadian Institute of Health Information (CIHI) has developed a new methodology [1,3] that adds a complexity overlay to most CMGs and makes another age adjustment for all CMGs and complexity levels if appropriate.

The complexity overlay developed by the CIHI divides the cases assigned to each CMG into four levels. The assignment of complexity level is based on the co-morbidities and complications. The complexity level can then be represented as follows:

1. no complexity;
2. complexity related to chronic condition(s);
3. complexity related to serious/important condition(s);
4. complexity related to potentially life-threatening condition(s).

The complexity overlay is meant to improve estimation of ELOS. A high complexity level reflects an expectation of prolonged stay and/or a higher demand on the resources of the hospital.

The age adjustment to CMGs are made in one of two ways, by categorizing patient age into three age groups (0–17, 18–69, 70 +) or by simply using age as a continuous variable if such is necessary or applicable to the predictive model used.

CIHI estimates the expected length of stay (ELOS) using the following regression model which incorporates medical complexity and age:

$$\text{ELOS} = \beta_0 + \beta_1\ \text{Complexity} + \beta_2\ \text{Age} + \beta_3\ \text{Complexity} \times \text{Age}. \tag{1}$$

In this model it is obvious that the LOS is assumed to follow a normal distribution. In several studies, ANOVA and MANOVA have been used to analyse the relation between hospital LOS or intensive care unit LOS and factors such as age, complexity scores. All these analyses must make the assumption that the LOS distributions are normal.[4–8]

However, the fact that LOS distributions are not necessarily normal and usually skewed has been shown repeatedly, shedding significant doubt on the validity of such predictive models. For example, Chu[9] has found that there were more than 50 per cent of the cases above the mean LOS in some high-volume CMGs. In addition, some LOS data should be considered censored, such as

LOS for patients who die, are transferred to another hospital, or sign out against medical advice. Fenn *et al.*[10] raised the concerns about the analysis of censored treatment cost data in economic evaluations and suggested applying the Kaplan–Meier estimator in analyses.

The author proposed the application of lifetime models to LOS analyses and used the traditional methods, such as logrank test and Wilcoxon test, to compare the overall cumulative discharge rate curves among different time periods or groups, compare the cumulative discharge rates at each time interval and estimate the instantaneous forces of discharge at each time interval.[11] In this paper the author applies lifetime models to the estimation of adjusted expected length of stay (AELOS).

2. METHODS

2.1. Stay Distribution Function

In order to investigate the distribution of length of stay, we first need to define the *stay distribution function*. Let T be a non-negative random variable representing the discharge time of an individual patient from a homogeneous population. The stay distribution function, which is similar to the survival function, could be defined as[12,16–19]

$$S(t) = \Pr(T > t). \tag{2}$$

Thus, the expected length of stay could be defined as

$$e = E(T) = \int_0^{\Omega} S(t)\, \mathrm{d}t \tag{3}$$

where Ω is the maximum discharge time. Considering that the time which the patient spends in the hospital is counted as days, we could designate time by t_i for the ith day ($i = 1,2, \ldots, k$), where $0 = t_1 < t_2 < \ldots < t_k = \Omega$.

Using the trapezoidal rule, we have the estimate of ELOS as follows

$$\hat{e} = \frac{1}{2} \sum_{i=1}^{k-1} [S(t_i) + S(t_{i+1})]\, b_i \tag{4}$$

where

$$b_i = t_{i+1} - t_i, \quad t_k = \Omega.$$

We can also define the discharge-force function $h(t)$, analogous to the hazard function,[12,16–18] which refers to the instantaneous probability of being discharged at time t given the patient stays up to time t. We have

$$h(t) = \lim_{\Delta \to 0^+} \frac{\Pr(t \leqslant T \leqslant t + \Delta \,|\, t \leqslant T)}{\Delta} \tag{5}$$

and

$$h(t) = f(t)/S(t) \tag{6}$$

where

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}.$$

2.2. Adjusted Stay Function

To investigate the specific stay distribution under a specific combination of factors, such as complexity and age, we could further define the adjusted stay function as

$$S(t \mid \boldsymbol{Z}) = \Pr(T > t \mid \boldsymbol{Z}) \tag{7}$$

where $\boldsymbol{Z}$ is the vector of factors which influence the length of stay.

Thus, making factor adjustment, we have the adjusted expected length of stay (AELOS)

$$e(\boldsymbol{Z}) = E(T \mid \boldsymbol{Z}) = \int_0^{\Omega} \mathrm{S}(t \mid \boldsymbol{Z})\, \mathrm{d}t. \tag{8}$$

When the factors are fixed (for example, complexity level = 3 and age = 40 years), the AELOS could be estimated from (8) once the $S(t \mid \boldsymbol{Z})$ is estimated.

2.3. Estimation of Adjusted Stay Function

The covariate adjustment of the stay function could be done based on various types of lifetime models (parametric, semi-parametric or non-parametric). This paper will discuss two types of models, the proportional hazards model and the Kaplan–Meier adjustment.

2.3.1. Proportional hazards model

The Cox proportional hazards regression model links the length of stay at an event ($T > 0$) to a set of covariates, specified by the row factor vector $\boldsymbol{Z} = (z_1, \ldots, z_k)$. This model has a property, referred to as the proportional hazards assumption, that different individuals have hazard functions which are proportional to one another. That is, the ratio $h(t \mid \boldsymbol{Z}_1)/h(t \mid \boldsymbol{Z}_2)$ of the hazard functions for two individuals with regression vectors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ does not vary with t.[16]

Therefore, the discharge-force function $h(t \mid z)$,[12,14,18] the instantaneous conditional probability that patients with factor vector $\boldsymbol{Z}$ are discharged at time t, could be expressed as

$$h(t \mid \boldsymbol{Z}) = h_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}) \tag{9}$$

where $h_0(t)$ is an arbitrary unspecified baseline discharge-force function and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ is a vector of unknown regression coefficients.

From equation (6), the model could be also expressed as

$$S(t \mid \boldsymbol{Z}) = [S_0(t)]^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})} \tag{10}$$

where $S_0(t) = \exp(-\int^t h_0(u)\, \mathrm{d}u)$ is the baseline stay function.

In our case, the clinical practice shows that the discharge-force is usually related with age quadratically and complexity linearly. It shows the discharge-force is related with the interaction of complexity and age as well. Thus, we could use this model to estimate the stay function for a particular combination of complexity and age by

$$S(t \mid \text{Complexity, Age}) = [S_0(t)]^{\exp(\beta_0 + \beta_1 \text{ Complexity} + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \beta_4 \text{ Complexity} \times \text{Age})}. \tag{11}$$

The proportional hazards assumption on $\boldsymbol{Z}$ needs to be checked. A stratified proportional hazards model may be more appropriate since complexity is defined categorically. Defining

$S_i(t \mid \text{Age})$ as the stay function for the ith complexity stratum, Cox proportional hazards models are fitted to age within each complexity stratum and age-adjusted discharge curves for each complexity could be presented as well. The stay function could be presented as

$$S_i(t \mid \text{Age}) = [S_{0i}(t)]^{\exp(\beta_{0i} + \beta_{1i}\text{Age} + \beta_{2i}\text{Age}^2)}. \tag{12}$$

2.3.2. Kaplan–Meier adjustment

While the stratified proportional hazards model avoids the proportional hazards assumption for complexity, it still makes the assumption for age. If neither age nor complexity meets the proportional assumption, a non-parametric approach such as the Kaplan–Meier age adjustment[15,20] is preferable.

Based on the specific situation, ages could be broken down into several groups. Then, the age-adjusted stay function for the ith age group and jth complexity could be obtained by

$$S_{ij}(t) = \frac{\sum_{k=1}^{P_{ij}} W_{ijk} S_{ijk}(t)}{\sum_{k=1}^{P_{ij}} W_{ijk}} \tag{13}$$

where $S_{ijk}(t)$ is the Kaplan–Meier estimate of the age specific stay function for the ith age group and the jth complexity group, P_{ij} is the number of various ages or various subgroups of ages occurring in the ith age group under the jth complexity and W_{ijk} is the number of patients at each age or in each subgroup of age at time 0.

3. AN EXAMPLE

The above approach was applied to a data set of lengths of stay of the patients, whose CMGs are CMG 179, coronary bypass with pump, no cardiac catheterization, with a variety of ages and complexities.

Using the Cox proportional hazards model and best subset selection, the complexity and age were found to be the significant factors influencing the length of stay ($\beta_{\text{Complexity}} = -0{\cdot}8251$, $\beta_{\text{Age}} = -0{\cdot}0434$, $p = 0{\cdot}0001$). The stay functions for the combinations of complexity and age were estimated and the expected lengths of stay with complexity and age adjustment were finally obtained (method 1). The results are shown in Table I.

Considering complexity as a categorical variable, we investigated the role of age under separate complexity levels. Using the Cox proportional hazards model, age was separately found to be a significant factor influencing the length of stay under all four complexity levels (complexity = 1, $\beta_{\text{Age}} = -0{\cdot}0418$, $p = 0{\cdot}004$; complexity = 2, $\beta_{\text{Age}} = -0{\cdot}0350$, $p = 0{\cdot}0001$; complexity = 3, $\beta_{\text{Age}} = -0{\cdot}0914$, $p = 0{\cdot}0001$; complexity = 4, $\beta_{\text{Age}} = -0{\cdot}0377$, $p = 0{\cdot}0097$). The stay functions for age under four complexity levels were estimated. Based on these results, the expected lengths of stay with age adjustment under different complexity levels were obtained (method 2). The results are also shown in Table I.

The validity of the proportional hazards assumption was checked. There was no evidence of increasing or decreasing trends over time in the hazard ratio for the models in method 2, while there was in method 1.

The martingale residual and deviance residual [21] were also used to investigate the lack of fit of these models separately. The results showed that there were no indications of a lack of fit of these models in method 2, while there was in method 1.

Table I. Expected length of stay (days) estimated by method 1 and method 2 based on the Cox proportional hazards model*

Age	Complexity level							
	1		2		3		4	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
45	4·22	4·35	5·44	5·15	7·91	6·66	11·34	11·98
50	4·46	4·62	5·95	5·54	8·75	8·07	12·39	13·36
55	4·75	4·90	6·57	6·05	9·63	9·58	13·56	14·52
60	5·10	5·21	7·28	6·69	10·56	11·01	14·88	15·49
65	5·53	5·54	8·07	7·48	11·54	12·30	16·34	16·39
70	6·07	5·88	8·92	8·39	12·61	13·45	17·91	17·44
75	6·71	6·23	9·82	9·41	13·81	14·46	19·54	18·99

* The estimates of the AELOS were based on the continuous ages. However, only several estimates are listed for the specific ages in the table because of space limitation

Table II. Expected length of stay estimated by Kaplan–Meier adjustment

Age group	Complexity level			
	1	2	3	4
45–49	4·51	5·54	6·50	13·50
50–54	4·71	6·28	7·87	14·61
55–59	4·96	6·91	9·02	15·62
60–64	5·20	7·62	10·79	16·67
65–69	5·60	8·11	12·31	17·50
70–74	6·67	10·35	14·38	19·47

This suggested that a stratified proportional hazards model is more appropriate than a non-stratified model for this study.

The Kaplan–Meier adjustment was also employed to estimate the adjusted stay functions for the specific combination of age group and complexity. Based on the results from these estimates of the adjusted stay function, the expected lengths to stay for specific age groups and complexities could be obtained. The results are listed in Table II.

4. DISCUSSION

A better predictor for length of stay assists with programme planning and evaluation, physician impact analysis, monitoring clinical practice, developing benchmarks, and discharge planning in the management of hospital resource utilization. An improved predictor for length of stay should be made based on the precise investigation of the distribution of length of stay.

In recent years concerns have risen about the fact that the distribution of length of stay is usually skewed and/or lepto-kurtotic rather than normal.[9,10] The use of lifetime models provides a powerful approach to investigate the distribution of length of stay. These approaches, which

include parametric, semi-parametric and non-parametric methods, are able to deal with the data whose distributions do not meet the assumption of normality and with censored data.

The lengths of stay for death, transfer and sign out are actually incomplete. The censoring mechanism in survival analysis can deal with such cases where subjects withdraw from a study or the study is completed before the endpoint is reached. In these cases, the survival times (also known as failure times) are censored; subjects survived to a certain time beyond which their status is unknown.[12,17,18] In this application, the stay times or discharge times for death, transfer and sign out could be considered censored. However, a further study would be conducted on the stay times for death, since death is related to complexity and age.

The Cox proportional hazards model, which is usually employed to analyse the relationship between survival time and risk covariates, could be also employed in the analysis of the length of stay. There are several approaches to adjust for complexity and age. One method is to use one model that includes the interaction of complexity and age. Another method is to use several models under several specific complexity levels. The latter one is more preferable because the complexity is categorical.

We can readily estimate the adjusted stay function using statistical software packages. Based on the results of the adjusted stay function, we can easily obtain the estimate of the expected length of stay with complexity and age adjustment, which gives us a graphic representation of the discharge process and helps to compare different discharge processes with complexity and age adjustment.

In situations that violate the proportional hazards assumption, the Kaplan–Meier method could provide more reasonable estimations of the stay function for a combination of age groups and complexity. They may also be estimated using various statistical software packages to obtain the ELOS using the same method as described in equation (3).

In summary, we have described two methods, semi-parametric and non-parametric, that produce the adjusted stay function using lifetime models. These methods are appropriate for use given the reality of the distribution of length of stay of patients in hospital.

In the future, other models, such as Weibull, exponential, gamma, and log-normal models, could be applied in the estimation of the stay function. The Poisson, overdispersed Poisson and negative binomial regression models could also be applied to deal with the discrete length of stay.

ACKNOWLEDGEMENTS

The author is grateful to Dr. M. Cusimano, Dr. S. Hwang and Dr. I. Fong for their helpful advice.

REFERENCES

1. Canadian Institute for Health Information. 'Complexity and age adjustment, 1995: expanding the CMG methodology', Canadian Institute for Health Information, 1995.
2. Pink, G. H. and Bolley, H. B. 'Physicians in health management: 3. Case Mix groups and Resource Intensity weights: an overview for physicians', *Canadian Medical Association Journal*, **150**, 889–894 (1994).
3. Canadian Institute for Health Information. 'DAD resource indicators for use with complexity 1997', Canadian Institute for Health Information, 1997.
4. Barie, P. S., Hydo, L. F. and Fisher, E. 'Utility of illness severity scoring for prediction of prolonged surgical care', *Journal of Trauma*: *Injury*, *Infection and Critical Care*, **40**, 513–519 (1996).
5. Becker, R. B., Zimmerman, J. E., Knaus, W. A., Wagner, D. P., Draper, E. A., Higgins, T. L., Estafanous, F. G. and Loop, F. D. 'The use of APACHE III to evaluate ICU length of stay, resource use and mortality after coronary artery by-pass surgery', *Journal of Cardiovascular Surgery*, **36**, 1–11 (1995).

6. Bertozzi, B., Barbisoni, P., Franzoni, S., Rozzimi, R., Frisoni, G. B. and Trabucchi, M. 'Factors related to length of stay in a geriatric evaluation and rehabilitation unit', *Aging*, **8**, 170–175 (1996).
7. Cullen, D. J., Apolone, G., Greenfield, S. Guadagnoli, E. and Cleary, P. 'ASA physical status and age predict morbidity after three surgical procedures', *Annals of Surgery*, **220**, 3–9 (1994).
8. Shabot, M. M. and Johnson, C. L. 'Outcome from critical care in the 'oldest old' trauma patients', *Journal of Trauma*: *Injury*, *Infection and Critical Care*, **39**, 254–260 (1995).
9. Chu, C. 'Resource intencity weighing and case mix grouping: assumptions and implications for health service performance evaluation', *Healthcare Management Forum*, **7**, 24–31 (1994).
10. Fenn, P., McGuire, A., Phillips, V., Backhouse, M. and Jones, D. 'The analysis of censored treatment cost data in economic evaluation', *Medical Care*, **33**, 851–863 (1995).
11. Li, J. L. 'An approach to evaluate efficiency of hospital resource utilization - comparison in length of stay', Proceedings of Association of the Management and the International Association of Management 14th Annual International Conference, 1996.
12. Cox, D. R. and Oakes, D. *Analysis of Survival Data*, Chapman and Hall, London, 1980.
13. Chiang, C. L. *The Life Table and its Applications*, Makabar, Robert E. Krieger Publishing Company, Florida, 1984.
14. Cox, D. R. 'Regression models and life tables', *Journal of the Royal Statistical Society*, *Series B*, **34**, 187–220 (1972).
15. Cupples, L. A., Gagnon, D. R., Ramaswamy, R. and D'Agostino, R. B. 'Age-adjusted survival curves with application in the Framingham study', *Statistics in Medicine*, **14**, 1731–1744 (1995).
16. Elandt-Johnson, R. C. and Johnson, N. L. *Survival Models and Data Analysis*, Wiley, New York, 1980.
17. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
18. Lawless, J. E. *Statistical Models and Methods for Lifetime Data*, Wiley, New York, 1982.
19. Lee, J., Yoshizawa, C., Wilkens, L. and Lee, H. P. 'Covariance adjustment of survival curves based on Cox's proportional hazards regression model', *Computer Applications in the Biosciences*, **8**, 23–27 (1992).
20. Kaplan, E. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
21. Themeau, T. M., Grambsch, P. M. and Fleming, T. R. 'Martingale-based residuals and survival models', *Biometrika*, **77**, 147–160 (1990).

STATISTICS IN MEDICINE
Statist. Med. **18**, 3345–3354 (1999)

MULTI-CRITERIA DECISION MAKING – AN APPROACH TO SETTING PRIORITIES IN HEALTH CARE

FLÁVIO FONSECA NOBRE[1]*, LILIAN TEREZINHA FERREIRA TROTTA[1] AND LUIZ FLÁVIO AUTRAN MONTEIRO GOMES[2]

[1] *Programa de Engenharia Biomédica - COPPE/UFRJ, Cidade Universitária - Ilha do Fundão, P.O. Box 68510, 21945-970 - Rio de Janeiro - RJ, Brazil*

[2] *Departamento de Engenharia de Produção, Universidade Federal Fluminense, 24210-240 - Niteroi - RJ, Brazil*

SUMMARY

The objective of this paper is to present a multi-criteria decision making (MCDM) approach to support public health decision making that takes into consideration the fuzziness of the decision goals and the behavioural aspect of the decision maker. The approach is used to analyse the process of health technology procurement in a University Hospital in Rio de Janeiro, Brazil. The method, known as TODIM, relies on evaluating alternatives with a set of decision criteria assessed using an ordinal scale. Fuzziness in generating criteria scores and weights or conflicts caused by dealing with different viewpoints of a group of decision makers (DMs) are solved using fuzzy set aggregation rules. The results suggested that MCDM models, incorporating fuzzy set approaches, should form a set of tools for public health decision making analysis, particularly when there are polarized opinions and conflicting objectives from the DM group. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

With traditionally limited resources, particularly in developing countries, health managers are usually faced with the problem of making rational choices among competing proposals. To assess the usefulness of competing projects, health decision makers (DM) generally rely on some quantitative decision analysis model, derived from a variety of disciplines, including operations research, statistics, economics and psychology.[1,2] The aim of the models is to guide the DM to compare alternatives and help in the selection of the most preferred solution for the decision situation. The range of approaches varies from the simplest one, based only on the intuition and expertise of the DM, to those involving a greater mathematical formalism, such as a linear programming model.

Two well known methods are cost-effectiveness analysis and cost-utility analysis. Both methods require DM to identify, measure and compare costs and desirable consequences or

* Correspondence to: Flávio Fonseca Nobre, Programa de Engenharia Biomédica - COPPE/UFRJ, Cidade Universitária, Ilha do Fundão, P.O. Box 68510, 21945-970 - Rio de Janeiro - RJ, Brazil, E-mail: flavio@peb.ufrj.br

Contract/grant sponsor: Brazilian Ministry of Science and Technology
Contract/grant number: PRONEX 41.96.0937.00, CNPq 521444/96-9

Contract/grant sponsor: CAPES/Brazilian Ministry of Education

CCC 0277–6715/99/233345–10$17.50

outcomes of available alternatives of addressing the decision problem. For effectiveness different units are used, which makes it impossible to compare available alternatives, specifically when we are comparing technologies for prevention, diagnosis and therapy. Utility usually involves the quality-adjusted life years (QALY), which is a measure of health-related utility calculated as life expectancy weighted for quality of life. Despite suggestions that this form of analysis may be useful for helping the DM, QALY is quite difficult to obtain, particularly in developing countries. Another approach, known as multiple- criteria decision making (MCDM), focuses on capturing the preference of decision makers instead of solving well-structured optimization problems.[3]

A well-known MCDM method is the analytic hierarchy process (AHP) introduced by Saaty,[4] that uses a so-called additive value function to capture the preferences of decision makers. One characteristic of this method is the phenomenon of rank reversal, that is, relative rankings of two alternatives can change depending on what other alternatives are considered.[5] Korhonen *et al.*[3] and Stewart[5] provide a good review of other MCDM that evolved using the utility function as a basic principle. However, although this has been the most used approach, utility theory has some limitations leading to inconsistencies in the ranking of alternatives.[6–8] An alternative, proposed by Kahneman and Tversky,[6] is the use of prospect theory to measure subjective criteria in explaining decision making behaviour.

This paper considers a multi-criteria method using prospect theory, in place of utility function, for describing the decision-maker evaluation of each criterion.[7,8] When more than one DM is involved in evaluating competing alternatives, aggregation methods must be used to combine the different viewpoints.[9,10] Therefore, it is also shown that fuzzy set aggregation methods can be valuable tools for ranking the alternatives. The applicability of the method is demonstrated using a decision problem related with purchasing of health technology for a Brazilian hospital.

2. MATERIALS AND METHODS

During the beginning of 1996, Pedro Ernesto Hospital, at Rio de Janeiro city in Brazil, was involved with the process of selecting technologies to be purchased. Pedro Ernesto Hospital is a university hospital with 600 beds, 1500 outpatients/day and 1200 inpatients/month. It includes most types of clinical and surgical departments and it is a reference centre for haemodialysis, computer tomography and nuclear medicine. About 60 per cent of all cardiac surgery in Rio de Janeiro State is done in this hospital. All professionals selected to participate in this process were medical doctors and two have experience in acquisition of equipment.

Eight technologies are identified as candidates for purchasing by the hospital, and all are recognized as established procedures. These technologies are: (i) magnetic resonance image (MRI); (ii) extracorporeal shock wave lithotripsor; (iii) single positron emission computer tomography (SPECT); (iv) video-laparoscope; (v) mamograph; (vi) fluorescein angiograph; (vii) cardio-angiograph; (viii) helical computed tomograph.

2.1. Multi-criteria Decision Making

Multi-criteria decision making comprises a broad set of methods and models that helps and guides the decision maker in discovering his or her most desired solution to the problem. The DM is interested in either choosing one alternative (a choice problem) or obtaining an order of preference of the alternatives (a ranking problem).[3–5, 8–11]

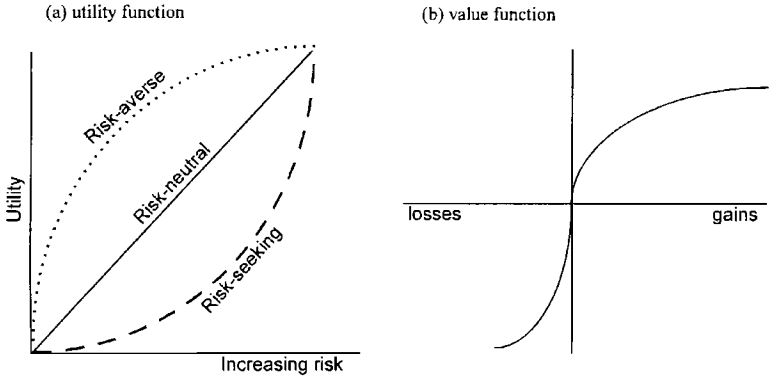


Figure 1. Typical functions of (*a*) utility theory and (*b*) prospect theory

A decision problem is based on a set of finite alternatives $A: \{a, b, c, \dots\}$ and a set of judgement criteria $C: \{1, \dots, k, \dots p\}$. Given the alternatives and criteria, a matrix $V: \{V_k(a)\}$ is obtained reflecting the viewpoint of the DM, where $V_k(a)$ corresponds to the evaluation of alternative $a\,(a \in A)$ with respect to criteria k. Value judgement for each criterion can be expressed on either a cardinal or a verbal scale. These scales are employed for ordering alternatives with respect to criteria and to weight the criteria. By directly using verbal scales, judgemental statements are converted into numerical values read on the cardinal scale.

The classical approach adopted in MCDM is to use a utility function. The shape of this function is concave-down for risk averse persons, suggesting that after a certain value of risk the utility grows slower than expected. For risk seeking behaviour, there is a tendency to increase utility faster with increased risk, resulting in a concave-up form. Neutrality is represented as a linear function (Figure 1(*a*)). To get a global evaluation and a ranking of the alternatives, the value judgement matrix is normalized, say to 0-1, for each alternative and transformed using a set of weights, g_k, associated with the importance of the criterion k. The final value of each alternative is obtained by summation of the transformed scores.

Like the utility function, for prospect theory the shape of the value function explains risk aversion and risk seeking behaviours. While utility expresses the psychological preferences of a DM, the value function is measuring the differences in utility or preference. For prospect theory, the proposed function of Kahneman and Tversky[6] is displayed in Figure 1(*b*). This function has an S-shape form that is concave above the reference or neutral point representing gains, and convex below the reference point representing losses. The concave part reflects risk aversion faced with gains, and the convex part is related with risk seeking behaviour when the DM is faced with losses. Because responses to losses are more extreme than to gains,[6] the slope of the function for losses is steeper.

Based on the prospect theory, an interactive multi-criteria decision making method (known as TODIM, Tomada de Decisão Interativa Multicritério in Portuguese), has been proposed by Gomes.[8] In this method, the matrix V, representing the value judgement of the DM, is normalized across alternatives obtaining a matrix p criteria $\times\, n$ alternatives' $W: \{w_c\,(a)\}$, which is called the position matrix of the DM following Znotinas and Hipel[9] terminology.

To use the proposed multi-criteria model, eight criteria are identified using a hierarchical approach, starting with a general goal and identifying sub-goals down the hierarchy that are more easily subject to evaluation by the DM. The criteria are: (i) benefit population; (ii) social impact; (iii) availability of human resources; (iv) dependence of facilities; (v) dependence on maintenance; (vi) community and professional demand; (vii) importance for improving patient condition; (viii) expected benefits in health outcomes. Of these, benefit population with the technology is the only one that is not qualitative and was based on the expected weekly number of exams for each technology.

The decision agents are requested to provide both the alternative weights with respect to each criterion and the criteria's weights using a semantic scale, resulting in a position matrix V for each DM. Also, they were asked to specify an ordered list of three alternatives they would suggest based on their expertise.

From matrix V, a position matrix W, representing the normalized value judgement of the DM, is obtained using a normalized set of weights, $G(g_c)$. The transformation is based on expressing the scores of each alternative as positive or negative deviations (gains or losses) from all other alternatives. For each DM and criterion an $n \times n$ matrix, $\mathbf{\Phi}_c$, called partial dominance is generated. The elements $\mathbf{\Phi}_c(i, j)$ are obtained using an additive difference function, that closely describes the shape of the prospect value function of Figure 1(*b*), and is given by the following expression:[8]

$$\Phi_c(i,j) = \begin{cases} \sqrt{\left\{\dfrac{g_c[w_c(i) - w_c(j)]}{\sum_C g_c}\right\}} & \text{if } w_c(i) - w_c(j) > 0 \\ 0 & \text{if } w_c(i) - w_c(j) = 0 \\ -\sqrt{\left\{\dfrac{(\sum_C g_c)(w_c(j) - w_c(i))}{g_c}\right\}} & \text{if } w_c(i) - w_c(j) < 0 \end{cases}$$

where, c is a generic criterion, g_c is the normalized weight of criterion c, and $w_c(i)$ and $w_c(j)$ are the normalized value judgements provided by the DM. It should be noted that $w_c(i) - w_c(j) > 0$ represents a relative gain, while $w_c(i) - w_c(j) < 0$ represents a relative loss.

TODIM can be used either for a single DM or for some group decision agents. In both cases, the analyst constructs an $n \times n$ dominance matrix $D: \{\delta(i, j)\}$, where n is the number of alternatives. The elements $\delta(i, j)$ are obtained using an additive function expressing the sum of the partial dominance for the attributes, and is in the form

$$\delta(i, j) = \sum_C \Phi_p(i, j), \quad \forall (i, j).$$

If $\delta(i, j) > 0$, then alternative i is said to dominate alternative j, that is, alternative i is preferable to alternative j; if $\delta(i, j) = 0$ alternatives i and j are equivalent. The overall values of the various alternatives are combined to produce a rank ordering by computing the following values:

$$\xi_i = \frac{\sum_{j=1}^{n} \delta(i,j) - \min_i \sum_{j=1}^{n} \delta(i, j)}{\max_i \sum_{j=1}^{n} \delta(i, j) - \min_i \sum_{j=1}^{n} \delta(i, j)}$$

where ξ_i is the overall value of alternative i.

For a single DM, the ranking derived from the last expression represents the solution of the MCDM. However, when more than one agent is involved in the decision process, the viewpoint of all DM must be considered in the final result. One alternative to aggregate different opinions is through fuzzy aggregation[9,10] that has its basis in fuzzy set theory. This allows us to deal with uncertainty and imprecision that are always present in decision makers' mind.

3. FUZZY SET AGGREGATION METHODS

As originally proposed by Znotinas and Hipel,[9] the set of criteria can be viewed as a fuzzy set F, and the DM is required to provide a value reflecting the degree of membership directly, using some qualitative scale. Here, the degree of membership is derived from each partial dominance matrix $\mathbf{\Phi}_c$ to obtain what is termed the net benefit evaluation matrix for position P of the DMs.

Therefore, a new position matrix P is derived by first computing for the DM and each criterion the value ξ_i, from the partial dominance matrix, substituting in the expression $\delta(i, j)$ for $\mathbf{\Phi}_c(i, j)$. The resulting expression is equivalent to a linear membership function. This new ξ_i express the degree of satisfaction of all alternatives for criterion c according to TODIM, and they represent degree of membership, p_i, evaluated from the DM. For this application four new position matrices are obtained.

In what follows we will use the structure and basic terminology presented in Znotinas and Hipel.[9] Given two fuzzy sets, F_1 and F_2, one may define the operations of union and intersection. For a given position matrix, representing the viewpoint of a DM, the evaluation of an alternative across the set of criteria is termed an *evaluation profile* for that fuzzy set. An evaluation profile for alternative a_n, position matrix P_m, for the mth DM, and criterion k is denoted by the fuzzy set $F_{anPm} = \{F_{aPm}(k)\}$. Considering the evaluation profile for alternative 1 and the four DMs, one can write these fuzzy sets as

$$F_{a1P1} = \{0{\cdot}27, 0{\cdot}00, 0{\cdot}00, 0{\cdot}00, 1{\cdot}00, 1{\cdot}00, 0{\cdot}00, 1{\cdot}00\}$$

$$F_{a1P2} = \{0{\cdot}26, 0{\cdot}00, 0{\cdot}00, 0{\cdot}00, 0{\cdot}78, 1{\cdot}00, 1{\cdot}00, 0{\cdot}00\}$$

$$F_{a1P3} = \{0{\cdot}26, 0{\cdot}00, 0{\cdot}00, 0{\cdot}26, 1{\cdot}00, 0{\cdot}68, 0{\cdot}85, 0{\cdot}74\}$$

$$F_{a1P4} = \{0{\cdot}27, 0{\cdot}00, 0{\cdot}00, 0{\cdot}00, 1{\cdot}00, 0{\cdot}00, 0{\cdot}00, 1{\cdot}00\}.$$

The union operation is the maximum of the respective elements from all four fuzzy sets and is given as

$$F_{a1P1} \cup F_{a1P2} \cup F_{a1P3} \cup F_{a1P4} = \max(F_{a1P1}, F_{a1P2}, F_{a1P3}, F_{a1P4})$$
$$= \{0{\cdot}27, 0{\cdot}00, 0{\cdot}00, 0{\cdot}26, 1{\cdot}00, 1{\cdot}00, 1{\cdot}00, 1{\cdot}00\}.$$

The intersection operation is given by the minimum of these elements with respect to each criterion, and is given as

$$F_{a1P1} \cap F_{a1P2} \cap F_{a1P3} \cap F_{a1P4} = \min(F_{a1P1}, F_{a1P2}, F_{a1P3}, F_{a1P4})$$
$$= \{0{\cdot}26, 0{\cdot}00, 0{\cdot}00, 0{\cdot}00, 0{\cdot}78, 0{\cdot}00, 0{\cdot}00, 0{\cdot}00\}.$$

Although other operations to obtain aggregate values from the evaluation matrices are available, these two operations are sufficient to illustrate the method used in this paper.

3.1. Aggregation of viewpoints

Znotinas and Hipel[9] suggested a number of alternatives to aggregate different viewpoints. Here we will describe only those that will be used to aggregate the matrices obtained from the decision agents involved in the priorization of technologies, that is: (i) the pessimistic aggregation; (ii) the mean aggregation, and (iii) the modified pessimistic aggregation.

(i) The pessimistic aggregation considers the 'worst' viewpoints in an attempt to minimize risk. It is obtained by using the intersection operation. This form of aggregation is used in most situations.[9]
(ii) The mean aggregation is simply an averaging calculation. The matrix is calculated by the arithmetic mean among the values of alternatives allocated by all decision makers for each criterion.
(iii) Finally, the modified pessimistic aggregation is obtained by averaging the mean aggregation with the pessimistic aggregation. The method tries to nullify the effects of extreme responses while maintaining the risk minimization characteristic of the pessimistic aggregation.[10]

3.2. Ranking of alternatives

For a given criterion k, an alternative j dominates another alternative i if its aggregate degree of membership is greater, that is $F_{aj\mathscr{P}}(k) > F_{ai\mathscr{P}}(k)$, where $\mathscr{P}$ is the aggregated matrix. Therefore, alternative j is superior to the alternative i if the evaluation profile of the first alternative dominates in more criteria the second alternative. To depict the dominance among all possible pair of alternative, Alley *et al.*[10] suggested the use of an $n \times n$ dominance matrix $D: \{d_{ij}\}$, where d_{ij} express for each pair of alternatives the number of attributes for which the membership value of alternative j (columns) dominates the alternative i (rows). The number of times an alternative j dominates all the others, is obtained by summing the jth column of the matrix. Similarly, the jth row total displays the number of times alternative j is dominated by all the other alternatives.

The process of ranking the alternatives starts by first comparing the column totals and identifying the two largest sums. Dominance is evaluated comparing the row totals of these two alternatives. An alternative completely dominates when its column sum is greater and its row sum is smaller than the other alternative. When column and row totals are greater, then discrimination between the two alternatives is obtained either if there is a larger difference in the dominance axes, or by examining the original position matrix. Once two alternatives are evaluated, the rows and columns associated with the alternatives are ignored and new column and row totals are calculated for evaluating the remaining alternatives. To show the usefulness of the approach, the results of the dominance matrix obtained from the modified pessimistic aggregation matrix, and corresponding analysis, are presented in the next section.

4. RESULTS

Figure 2 shows, for each of the four DMs, the overall values obtained for the TODIM multi-criteria approach making use of the original expression for (ξ_i). This figure is a convenient form of eliciting the preference structure of the DMs. A similar figure (not shown) is obtained when the criterion 'benefit population' is excluded from the analysis. It can be seen that alternatives 4, 5 and 7 are probably the most preferable for the group.

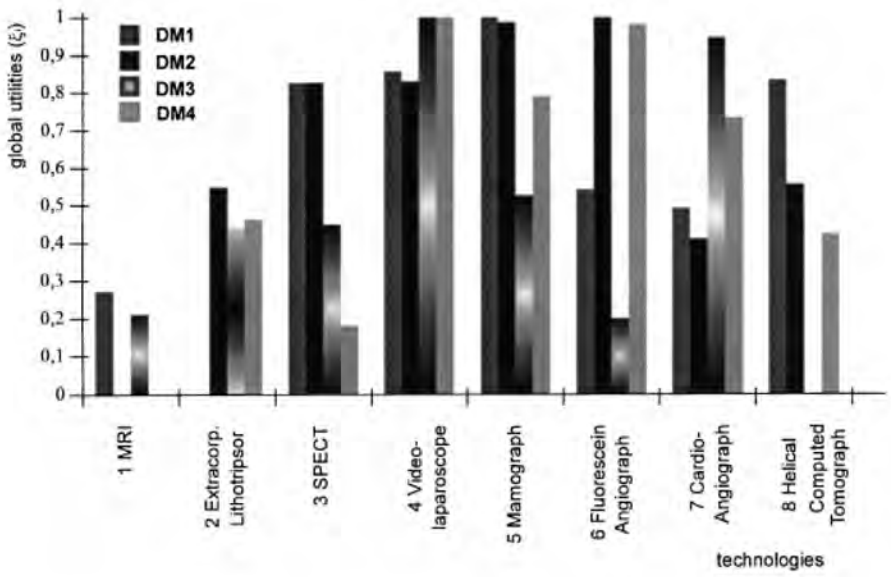


Figure 2. Priority scores according to each decision maker with the criterion 'benefit population with the technology'.

In this paper, only the modified pessimistic aggregate matrix is considered for construction of the dominance matrix and ranking the alternatives. The modified pessimistic aggregate matrix for the four DMs involved in this study, with the corresponding dominance matrix, is shown in Table I. The results of the dominance matrix clearly reinforce the dominance of alternatives 4 and 5 over the other alternatives. Comparing the column totals and respective row totals, there is an indication that alternative 4 completely dominates alternative 5. Ignoring the columns and rows of these alternatives in the dominance matrix, new column and row sums are obtained. As shown in Table I, alternatives 7 and 8 are the next choices. Because alternative 7 dominates in more criteria (26 versus 20), and is less dominated (14 versus 19), alternative 7 ranks higher than alternative 8. Computing new columns and rows sums, the dominance matrix shows that the next choices are alternatives 3 and 1. The remaining two technologies are evaluated, resulting in alternative 2 as the next 'best' choice and alternative 6 ranking as the last alternative.

To check the sensitivity of the results, a new dominance matrix has been obtained using weighted positions matrices.[9,10] These matrices are obtained by raising the values of each position matrix using as exponents the criteria's weight provided by the DM. This sensitivity analysis technique smooths the differences in the position matrix and could reduce possible bias in the original ratings. For this study, the weighted modified pessimistic aggregation approach produces similar ordering for the alternatives, showing a low sensitivity and reinforcing confidence in the findings.

5. DISCUSSION

In this investigation the TODIM methodology allowed the identification of potential differences between a decision based on the DM intuition and that considering explicitly a set of dimensions

Table I. Modified pessimistic aggregation matrix for all decision makers

Criteria	Alternatives							
	1	2	3	4	5	6	7	8
1	0·26	0·13	0·59	0·41	0·41	1·00	0·00	0·64
2	0·00	0·13	0·67	1·00	0·78	1·00	0·33	0·26
3	0·00	0·29	0·30	0·73	0·76	0·77	0·18	0·17
4	0·03	0·36	0·75	0·33	0·91	0·36	0·59	0·25
5	0·86	0·28	0·27	0·78	0·75	0·23	0·38	0·54
6	0·34	0·27	0·30	0·34	0·27	0·22	1·00	0·34
7	0·23	0·62	0·14	0·78	0·67	0·20	0·36	0·22
8	0·34	0·16	0·15	0·77	0·31	0·05	0·38	0·21

Dominance matrix

	1	2	3	4	5	6	7	8			
1	–	4	4	6	5	4	6	4	33	22	12
2	4	–	5	7	7	3	5	5	36	22	12
3	4	3	–	6	6	4	4	5	32	20	11
4	1	1	2	–	2	3	2	1	12		
5	3	0	2	5	–	3	2	2	17		
6	4	4	4	4	5	–	5	4	30	21	12
7	2	3	4	6	6	3	–	2	26	14	
8	3	3	3	6	6	4	6	–	31	19	13
	21	18	24	40	37	24	30	23			
	17	17	20			18	26	20			
	12	11	13			11					

of the decision problem. Examining the three listed alternatives for each DM (Table II), it is noted that for two DMs, two alternatives are included among the first three options from TODIM. For the remaining DMs only one is among the top three results of the TODIM approach. A possible explanation is that refining the decision domain in a formal set of criteria gives to the DM more knowledge of his/her preference structure and takes into account the inherent subjectivity of the human process for choices.

Using the modified pessimistic aggregation, which is appropriate when there are possibilities of polarized responses from the involved decision agents,[10] it was possible to find a group ranking structure that can help the practical setting of purchasing priorities. Comparing this ranking with the ordered list of preferences provided by the DMs, some interesting conclusions can be drawn. From Table II and the dominance matrix (Table I), the alternatives that ranked in the first three positions have larger dominance over the other available options. The resulting rankings of the alternatives suggest that the three most preferred technologies are alternatives 4, 5 and 7. In the provided preferred lists of the DMs, all three alternatives are mentioned. For the other alternatives, the major finding is that alternative 1, ranked in sixth place using the fuzzy aggregation approach, has been initially considered as first or second choice by three of the DMs.

While it is not possible directly to compare the ranking from the modified pessimistic dominance matrix to that derived using the averaged TODIM score (Table II), some basic

Table II. Global value ξ_i for the DMs. The prior DM preference is in brackets. Bold faces indicates the resulting first three alternatives from the TODIM method for each DM. Italics means least desirable results from TODIM. $R1$ is the rank from the averaged scores and $R2$ is the rank from the fuzzy set analysis approach

Alternative	DM1		DM2		DM3		DM4		$R1$	$R2$
1	0·27	(1)	*0·00*	(2)	0·21	(1)	*0·00*		8	6
2	*0·00*	(least)	0·55		0·45		0·46		7	7
3	0·83		**0·83**	(3)	0·45		0·18		5	5
4	**0·86**		**0·83**	(1)	**1·00**		**1·00**	(2)	1	1
5	**1·00**		**0·99**		**0·53**	(3)	**0·79**	(3)	2	2
6	0·54		**1·00**		0·20	(least)	**0·98**	(least)	3	8
7	0·49	(3)	0·41	(2)	**0·94**		0·73		4	3
8	**0·83**	(2)	0·56		*0·00*	(2)	0·42	(1)	6	4

comments are in order. Alternatives 4 and 5 are also classified as first and second preferred solutions. Alternative 7, which ranked third by fuzzy set approach, is classified as fourth. It can be noted that alternative 6, which was considered as the least desirable solution from the modified pessimistic dominance matrix, is the third ranked option for the mean TODIM score. This difference in ranking can be explained by observing that the dominance matrix approach discards alternatives as soon as they are classified. If we consider only the first column and row totals of the dominance matrix, alternative 6 is ranked as fourth, suggesting therefore that both processes produce similar results when considering all alternatives.

One important element for the appropriate use of the methods described here, particularly the fuzzy aggregation procedures, is the definition of the set of attributes or criteria and the design of the questionnaire. A possible methodological option for enhancing the selection and definition of the set of criteria is to use concepts of knowledge engineering approach. A study is planned to evaluate the application of this methodology in a variety of decision scenarios. In this paper we have presented the elements of an approach to support multi-criteria decision making in the context of health care. We suggest that such methodology could be carried out to give DMs confidence in selecting possible courses of actions, such as purchasing equipment or selecting alternative public health programmes.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian Ministry of Science and Technology (PRONEX research grant 41.96.0937.00 and CNPq research grant 521444/96-9). One of the authors, Lilian T. F. Trotta, was under a scholarship from CAPES/Brazilian Ministry of Education. Finally, thanks are also due to Dr. Roberto Magalhães, technical manager of Pedro Ernesto Hospital.

REFERENCES

1. Parker, B. R. 'In quest of useful health care decision models for developing countries', *European Journal of Operational Research*, **49**, 279–288 (1990).
2. Coast, J. 'The role of economic evaluation in setting priorities for elective surgery', *Health Policy*, **24**, 243–257 (1993)
3. Korhonen, P., Moskowitz, H. and Wallenius, J. 'Multiple criteria decision support - A review', *European Journal of Operational Research*, **63**, 361–375 (1992).

4. Saaty, T. L. *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, RWS Publications, Pittsburgh, 1994.
5. Stewart, T. J. 'A critical survey on the status of multiple criteria decision making theory and practice', *OMEGA International Journal of Management Science*, **20**, 569–586 (1992).
6. Kahneman, D. and Tversky, A. 'Prospect theory: an analysis of decision under risk', *Econometrica*, **47**, 263–291 (1979).
7. Hink, R. F. and Woods, D. L. 'How human process uncertain knowledge: an introduction for knowledge engineers', A*I Magazin*e, **Fall**, 41–53 (1987).
8. Gomes, L. F. A. M. and Lima, M. M. P. P. 'From modelling individual preferences to multicriteria ranking of discrete alternatives: a look at prospect theory and the additive difference model', *Foundations of Computing and Decision Sciences*, **17**, 172–184, (1992).
9. Znotinas, N. M. and Hipel, K. W. 'Comparison of alternative engineering designs', *Water Resources Bulletin*, **15**, 44–58 (1979).
10. Alley, H., Bacinello, C. P. and Hipel, K. W. 'Fuzzy set approaches to planning in the Grand River Basin', *Advances in Water Research*, **2**, 3–12 (1979).
11. Keeney, R. L. and Raiffa, H. *Decisions with Multiple Objectives*: *Preferences and Value Tradeoffs*, Wiley, New York, 1976.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3355–3363 (1999)

RECURRENT INJURY EVENT-TIME ANALYSIS†

JAMES T. WASSELL,[1]* WILLIAM C. WOJCIECHOWSKI[1‡] AND DEBORAH D. LANDEN[2]

[1] *Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Safety Research, 1095 Willowdale Road, Morgantown, WV 26505-2888, U.S.A.*
[2] *Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Pittsburgh Research Center, P.O. Box 18070, Pittsburgh, PA, 15236-0070, U.S.A.*

SUMMARY

Public health decision making based on data sources that are characterized by a lack of independence and other complicating factors requires the development of innovative statistical techniques. Studies of injuries in occupational cohorts require methods to account for recurrent injuries to workers over time and the temporary removal of workers from the 'risk set' while recuperating. In this study, the times until injury events are modelled in an occupational cohort of employees in a large power utility company where employees are susceptible to recurrent events. The injury history over a ten-year period is used to compare the hazards of specific jobs, adjusted for age when first hired, and race/ethnicity differences. Subject-specific random effects and multiple event-times are accommodated through the application of frailty models which characterize the dependence of recurrent events over time. The counting process formulation of the proportional hazards regression model is used to estimate the effects of covariates for subjects with discontinuous intervals of risk. In this application, subjects are not at risk of injury during recovery periods or other illness, changes in jobs, or other reasons. Previous applications of proportional hazards regression in frailty models have not needed to account for the changing composition of the risk set which is required to adequately model occupational injury data. Published in 1999 by John Wiley & Sons, Ltd. This article is a US Government work and is in the public domain in the United States.

INTRODUCTION

Data from occupational cohort studies may suffer from the same complications as other types of cohort or longitudinal studies. These include staggered accrual of subjects, loss to follow-up of subjects, changes in the composition of the risk-set due to changes in job status or recovery time from injury, changing values of important covariates over time and recurrent events. However, occupational injury data pose an additional problem requiring non-standard methods because individuals may have unobserved risk factors for injury. As a result, statistical models that account for unobserved effects provide substantial improvement compared to models that ignore individual random effects. The distribution of the unobservable heterogeneity is identifiable when there is a common subject-specific random effect for all the event-times recorded for one subject. Frailty models[1–7] can account for a complicated data structure and incorporate random effects to

* Correspondence to: James T. Wassell, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Safety Research, 1095 Willowdale Road, Morgantown, WV 26505–2888, U.S.A. E-mail: jtw2@cdc.gov

† This article is a US Government work and is in the public domain in the United States
‡ Currently at Rice University, Department of Statistics, Houston, TX 77251, U.S.A.

Published in 1999 by John Wiley & Sons, Ltd.

compare the injury hazards among different jobs using a method that is consistent with the natural history of occupational injuries.

Frailty models have been shown to be useful for analysis of correlated survival or event-time data. These methods propose that dependency between injury event-times for one individual is the result of an unobserved random variate that is common to all the injury event-times for one individual. This random variate has been termed 'frailty,' and acts as a multiplier which enhances or degrades the hazards for all times common to one individual. This approach has the advantage of modelling a possible 'mechanism' for dependent data as compared to other approaches[8–10] which adjust variance estimates to compensate for the effects of data dependency. Initially, frailty analyses were restricted to models that assume a parametric form for the 'baseline' (independent data models) survival time distributions. Parameters for these models may be estimated using traditional maximum likelihood methods, although the computing methods are complicated and rely on specialized computer programs.[11,12] Recent work has illustrated the use of the expectation maximization (EM) algorithm[13] and profile likelihood methods[14] to estimate parameters for frailty models not requiring parametric assumptions of the survival time distributions. In considering a counting process approach to frailty models, a recent paper by Nielsen *et al.*[15] describes frailty as 'accident proneness.' These models rely on iterative computing methods to obtain covariate estimates from Cox regression models assuming an arbitrary baseline hazard function in the presence of a frailty effect.

The distributional form of the random effects or frailties determines the structure and characteristics of the dependency among event-times. Several different distributions for the frailties have appeared in the literature;[16] the first proposed frailty model assumed that the frailties were distributed as gamma random variates with an expected value equal to one and a variance parameter that reflects the degree of dependency in the data.[17] Because the gamma and other frailty distributions suffer from a lack of the proportional hazards property in the marginal distribution, Hougaard[18] proposed the use of the positive stable frailty distribution. This analysis of recurrent injury data is based on the positive stable frailty model to take advantage of the proportional hazards property which allows interpretation of the effects of covariates in a straightforward manner.

For subjects who have recurrent events, the counting process formulation[1,19] of the proportional hazards model provides some advantages over the standard Cox regression analysis. For each subject an interval of risk is defined from the start of an at-risk period, to the time of an injury event or censored time. This approach has the advantage that the risk set for any event only includes other individuals whose interval of risk includes the event-time of interest. The method described here is intended to model events in calendar time through the use of the counting process formulation and also includes unobserved random variates (subject-specific frailty) to analyze multiple event data.

Cohort description

Ten years of injury records of employees of an electrical utility corporation who worked from 1980 to 1989 were analyzed in this study. The cohort consisted of 608 employees aged 18 to 24 years on 1 January 1980. Employees' entry into the study was staggered, although 84 per cent of the cohort were actively employed on 1 January 1980. Job titles included in the cohort were lineman, cable splicer, and others (electrician, troubleman, foreman and serviceman). Injuries were included in the study that occurred on the job and required the attention of a physician.

More than 400 employees experienced at least one injury and more than 250 employees experienced two or more injuries in this cohort. Following injury, an employee might not return to work for a period of recuperation or might return to work at a temporary office assignment until fully ready to resume former job duties. Employee records also indicated withdrawal from the workforce due to other reasons (for example, job change). The counting process model formulation was used to effectively account for these 'not at risk' periods.

Covariates investigated in this study included indicator variables for race: Black, Hispanic and White (referent group), and age at start of the study as a continuous covariate. The focus of the analysis is to compare three different job categories (linemen, cable splicers and others) regarding injury hazard. Additional details regarding this study can be found in an analysis of this data based on a nested case-control methodology[20] and an analysis based on the Weibull distribution for event-times with covariates and gamma frailty.[12]

METHODS

The following description of frailty models is adapted from Klein[13] and Wang *et al.*[14] Frailty models are based on the assumption that, if the value of the frailty were known, all the event-times for injuries occurring to an individual would be independent. Event-times, T_{ij}, are indexed by $i = 1, \ldots, B$, to indicate the ith subject with one or more recorded event-times denoted by the index $j = 1, 2, \ldots, n_i$. For each event time, a censoring indicator is recorded, I_{ij}, with $I_{ij} = 1$ if T_{ij} is an injury time and $I_{ij} = 0$ otherwise. Additional covariates may be included that are subject characteristics or specific to a subject at the time of an injury, $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijk})$ with k recorded covariates at T_{ij}. The unobserved frailty random variable, which varies among subjects but is common to all event-times for a subject, is denoted as W_i. Assuming a proportional hazards model, the independence assumption is demonstrated by defining the hazard rate as $\lambda(t_{ij}|\mathbf{Z_{ij}}, W_i) = W_i\lambda_0(t_{ij})\exp(\beta\mathbf{Z}_{ij})$ with $\lambda_0(t_{ij})$ as the arbitrary baseline hazard function and β is a vector of coefficients to be estimated. The joint survival function for all event-times recorded for the ith subject, given the frailty, is

$$P[T_{ij} > t_{ij}, j = 1, \ldots, n_i | W_i, \mathbf{Z}_{ij}, j = 1, \ldots, n_i] = \exp\left\{ -\left[\sum_{j=1}^{n_i} \Lambda_0(t_{ij})\exp(\beta Z_{ij})\right] W_i \right\}.$$

The discontinuous intervals of risk for subjects are accommodated by estimating the hazard function

$$\lambda_0(t_{ij}) = \frac{d_{ij}}{\sum_{t_{ij}} Y_i(t_{ij}) W_i \exp(\beta \boldsymbol{Z}_{ij})}$$

using an indicator function, $Y_i(t_{ij})$, to specify whether or not subject i is 'at risk' at time t_{ij}, where d_{ij} is the number of events that occur at time t_{ij}. The cumulative hazard, $\Lambda_0(t_{ij}) = \sum_{t \leqslant t_{ij}} \lambda_0(t_{ij})$ is a sum of the hazard values associated with all times less than or equal to t_{ij}.

Positive Stable Frailty Model Estimation

If a random variable has a positive stable distribution, $W \sim \boldsymbol{PS}(\rho)$, $W \geqslant 1$, indexed by the parameter $\rho \in (0, 1]$, then we can take advantage of the Laplace transform: $\boldsymbol{E}[\exp(-sW)] = \exp(-s^{\rho})$. When applied to the joint survival function above, substituting

$s = -[\sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\beta Z_{ij})]$ yields a joint unconditional survival probability suitable for constructing a log-likelihood function (shown in the Appendix) that can be used in parameter estimation. Note that independence of the event-times occurs if $\rho = 1$ and smaller values of ρ result in greater group dependency. Under independence, (that is, $\rho = 1$, so that all $w_i = 1$) the log-likelihood equation reduces to

$$\mathrm{LL}_{\text{independence}} = \sum_{j=1}^{n_i} I_{ij}[\beta Z_{ij} + \log[\lambda_0(t_{ij})]] - H_i$$

which provides the same estimates of regression coefficients as obtained from a Cox regression ignoring any grouping of the data.

To estimate parameters for a semi-parametric model with frailty an iterative method is required. Klein[13] proposed the EM algorithm approach to estimation for the gamma frailty proportional hazards regression analysis of family-grouped survival times. These analyses did not require the counting process approach for estimating covariate effects because family members (unlike members of an occupational cohort) do not leave the risk set for short periods of time.

In this study, we utilized a profile likelihood approach based on an EM algorithm for parameter estimation.[14] The profile likelihood requires determining the expected value of the frailties, given the current estimates of the regression parameters, the frailty parameter, and the baseline hazard rate. The profile likelihood estimation requires 'imputing' values for the individual frailties and using those estimates as an 'offset' term in a Cox regression estimation program. The estimation algorithm proceeds as follows:

1. Estimate the regression parameters under independence using the counting process formulation of the proportional hazards model.
2. Expectation step: calculate the expected value of the frailties given the regression estimates and the non-parametric estimate of the hazard (this hazard estimate depends on the current estimates of regression parameters).
3. Maximization step: use the group-specific estimates of the frailties as an 'offset' term in the proportional hazards model to update estimates of the regression parameters.
4. Iterate between steps 2 and 3 until parameter estimates converge, each time obtaining 'better' imputed values for the frailties.
5. Calculate the value of the log-likelihood that corresponds to the set of parameter estimates.
6. Find the set of parameter estimates that corresponds to the maximum of log-likelihood either graphically or with the use of a maximization routine.

Custom functions were written using S-plus statistical software[21] to perform the estimation for the positive stable frailty model. An optimization function (nlminb) was used to select the set of parameter estimates to maximize the profile likelihood. These functions (named 'frailty.stable' are available through StatLib (http://lib.stat.cmu.edu/S/).

Estimates of the standard error of the regression coefficients were obtained using the group jack-knife technique.[22,23] The group jack-knife estimates are obtained by multiple reanalysis of the data, each time omitting the group of injury records for one employee. The standard error estimates are equivalent to 'robust' estimates described elsewhere.[8] In addition, the group jack-knife estimates were compared to grouped bootstrap estimates based on a resampling of employee records, defining a group as all injury records for an employee. Hazard ratio estimates are not a linear combination of the coefficient estimates, therefore 95 per cent confidence intervals

for the hazard ratio estimates were obtained using the bias corrected accelerated percentile of the bootstrap replicates, rather than the estimated standard errors of the coefficients. A confidence interval for the frailty parameter was also obtained from the profile likelihood for comparison.

RESULTS

Parameter estimates and standard errors for the positive stable and for the independence model are shown in Table I. The estimated frailty parameter is 0·443 for the positive stable model. A comparison of the values of the log-likelihoods at its maximum for the frailty parameter ($-3445{\cdot}97$ at rho $= 0{\cdot}443$) and for the null model of no correlation ($-5558{\cdot}90$ at rho $= 1$) indicates that the model which accounts for the dependency in the data is significantly better. The profile log-likelihood 95 per cent confidence interval for the frailty parameter of the positive stable model is (0·420, 0·465) as indicated in Figure 1. The profile log-likelihood confidence interval is 23 per cent shorter than a 95 per cent confidence interval based on the group jack-knife standard error estimate of 0·015 (0·414, 0·472) and 26 per cent shorter than the BCa bootstrap confidence interval using percentiles of 250 bootstrap samples (0·416, 0·477).

With significant evidence of dependency in the data as indicated by the likelihood ratio tests of the frailty parameter, the interpretation of the covariate effects in the marginal distribution is affected. For example (under independence), a comparison of the hazard for the job title 'splicer' and the referent group (electrician, others) is 1·448 [$\exp(0{\cdot}370) = 1{\cdot}448$] with a 95 per cent BCa bootstrap confidence interval of (1·092, 1·820), suggesting this job has a significantly higher hazard of injury. Given subject-specific random effects, the hazard estimate based on the positive stable frailty model is attenuated to 1·050 [$\exp(0{\cdot}443 \times 0{\cdot}110) = 1{\cdot}050$] with a 95 per cent BCa bootstrap confidence interval of (0·950, 1·150), indicating that there is really no significantly higher hazard of injury for splicers. Table I illustrates other differences between the stable frailty model and hazard ratio estimates that ignore dependency; for example, the frailty model indicates that Hispanic workers have significantly higher injury hazard that is undetected by ignoring random effects.

DISCUSSION

Occupational cohort data on injury incidence and recurrence require analytical methods that account for the unique features of this data. In addition to some characteristics common to other longitudinal cohort studies, injury cohort data demonstrate the usefulness of random effect event-time models. The random effects account for the notions of unequal liability that have been long recognized by injury data investigators.[24] Frailty models account for an individual random effect that degrades or enhances the hazard function for injury for an individual. In contrast to modelling approaches that account for recurrent events through robust variance estimates, frailty models propose a mechanism for dependency in the data. In a cohort study, it is not possible to identify all risk factors for occupational injury, but additional unobserved factors influencing individual risk for injury may be accounted for by the use of frailty models.

An iterative method for estimation of parameters, using principles of the EM algorithm and based on a profile likelihood approach, has been developed. The algorithm iterates between expected values of the frailties, given regression estimates, and expected values of regression parameters, given frailty estimates, for a fixed value of the frailty distribution parameter. In addition to the counting process approach to analysis of multiple-event-time data, frailty models

Table I. Parameter estimates, standard errors, hazard ratios and 95 per cent confidence intervals obtained from the positive stable frailty model for correlated event-times and the Cox proportional hazards regression model for independent event-times

Effect	Positive stable frailty model				Cox regression model			
	Estimate	Std.Err.*	HR† = exp($\rho\beta$)	BCa 95% CI‡	Estimate	Std.Err.§	HR† = exp(β)	BCa 95% CI‡
Frailty (ρ)	0·443	0·015	—	(0·416, 0·477)	1·000	—	—	—
Black	0·076	0·092	1·034	(0·957, 1·119)	0·018	0·095	1·018	(0·845, 1·230)
Hispanic	0·288	0·135	1·136	(1·033, 1·265)	0·185	0·108	1·203	(0·963, 1·494)
Lineman	0·287	0·099	1·135	(1·031, 1·239)	0·408	0·112	1·503	(1·196, 1·804)
Splicer	0·110	0·110	1·050	(0·950, 1·150)	0·370	0·128	1·448	(1·092, 1·820)
Age	− 0·051	0·017	0·978	(0·964, 0·991)	− 0·039	0·017	0·961	(0·923, 0·994)

* Standard error estimates obtained from a group jack-knife procedure
† Hazard ratio point estimates
‡ Bias corrected accelerated percentile confidence intervals based on 250 bootstrap replicates
§ Standard error estimated from the information matrix

Published in 1999 by John Wiley & Sons, Ltd.

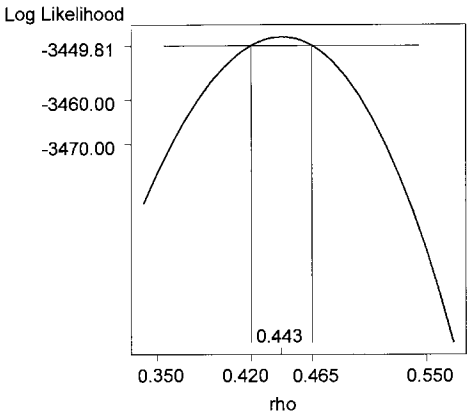


Figure 1. Log-likelihood profile confidence interval for the positive stable frailty distribution parameter (rho) based on a likelihood ratio test with 1 degree of freedom.. The log-likelihood function is maximum (-3445.97) at rho $= 0.443$. See the discussion for a comparison of this confidence interval with confidence intervals based on the group jack-knife variance estimate and bootstrap BCa confidence intervals

demonstrate additional improvement in estimating covariate effects. The additional complications of possibly censored times, variable 'at risk' periods and staggered accrual of subjects are accounted for in this analysis through the use of the counting process formulation of the proportional hazards model.

Additional research is needed regarding the appropriate choice of frailty distribution for a particular data set. Ideally, methods should be developed that permit the data to dictate the most suitable frailty distribution. This would provide valuable insight into the process generating data dependency in a given situation, similar to work which has focused on the gamma frailty models.[25] Further studies are needed to assess the sensitivity of estimation methods to the degree of correlation that may be present in realistic data. With significant frailty, the interpretation of the estimated coefficients depends on the choice of frailty distribution and the estimate of the frailty parameter so that the proportional hazards interpretation is possible only with the positive stable frailty model. Additional research is needed to develop methods to assess the validity of the proportional hazards assumption for the marginal distributions of multivariate failure time data. Although there is considerable computing effort required in fitting these models, these methods are becoming more widely available for data analysis.

APPENDIX

The log-likelihood for event-time data with positive stable frailty can be written as

$$\mathrm{LL} = \sum_{i=1}^{B} [D_i[\log(\rho) + (\rho - 1)\log(H_i)] - H_i^{\rho} + \log[\mathrm{J}(D_i, H_i)] + \sum_{j=1}^{n_i} I_{ij}[\boldsymbol{\beta Z_{ij}} + \log[\lambda_0(t_{ij})]]$$

where

$$J(D_i, H_i) = \sum_{m=0}^{D_i - 1} C_{D_i, m} H_i^{-m\rho}$$

based on a recursive function

$$C_{k,0} = 1$$

$$C_{k,m} = C_{k-1,m} + C_{k-1,m-1}[(k-1)\phi - (k-m)] \quad (m = 1, \ldots, k-2)$$

$$C_{k,k-1} = (\phi - 1)(2\phi - 1) \ldots [(k-1)\phi - 1] = \phi^{k-1}\Gamma(k-\rho)/\Gamma(1-\rho)$$

$$\phi = 1/\rho.$$

Also, $\lambda_0(t_{ij})$ is the hazard function at time t_{ij} described in the text. $H_i = \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{\beta Z_{ij}})$, for the cumulative hazard function $\Lambda_0(t_{ij})$ described in the text, and $D_i = \sum_j I_{ij}$ is the number of events that occurred to the ith subject, where there are a total of B subjects with n_i event-times for each subject. The expected value of the frailty, given current estimates of other parameters and the data is

$$E[W_i|\boldsymbol{\beta}, \lambda_0(t_{ij})] = \frac{E[W_i^{D_i+1}\exp(-H_iW_i)]}{E[W_i^{D_i}\exp(-H_iW_i)]}$$

$$E[W^q \exp(-H_iW)] = \rho H_i^{\rho-1} \exp(-H_i^{\rho})J(q, H_i), \quad q = 0, 1, \ldots; \quad H_i > 0$$

using the J function described above.

REFERENCES

1. Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer-Verlag, New York, 1993.
2. Costigan, T. M. and Klein, J. P. 'Multivariate survival analysis based on frailty models', *in* Basu, A. P. (ed.), *Advances in Reliability*, Elsevier Science Publishers B.V., Oxford, 1993, pp. 43–58.
3. Klein, J. P. and Moeschberger, M. L. *Survival Analysis*: *Techniques for Censored and Truncated Data*, Springer-Verlag, New York, 1997.
4. Oakes, D. 'Frailty models for multiple event times' *in* Klein, J. P. and Goel, P. K. (eds), *Survival Analysis*: *State of the Art*, Kluwer Academic Publishers, Dordrecht, 1992, pp. 371–379.
5. Pickles, A. and Crouchley, R. 'Generalizations and applications of frailty models for survival and event data', *Statistical Methods in Medical Research*, **3**, 263–278 (1994).
6. Wassell, J. T., Kulczycki, G. W. and Moyer, E. S. 'Modeling frailty in manufacturing processes' *in* Jewell, N. P., Kimber, A. C., Lee, M-L. T. and Whitmore, G. A. (eds), *Lifetime Data*: *Models in Reliability and Survival Analysis*, Kluwer Academic Publishers, Dordrecht, 1996, pp. 353–361.
7. Wassell, J. T., Kulczycki, G. W. and Moyer, E. S. 'Frailty models of manufacturing effects', *Lifetime Data Analysis*, **1**, 161–170 (1995).
8. Therneau, T. M. and Hamilton, S. A. 'rhDNase as an example of recurent event analysis', *Statistics in Medicine*, **16**, 2029–2047 (1997).
9. Lin, D. Y. 'Cox regression analysis of multivariate failure time data, the marginal approach', *Statistics in Medicine*, **13**, 2233–2247 (1994).
10. Li, Q. H. and Lagakos, S. W. 'Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event', *Statistics in Medicine*, **16**, 925–940 (1997).
11. Wassell, J. T. and Moeschberger, M. L. 'A bivariate survival model with modified gamma frailty for assessing the impact of interventions', *Statistics in Medicine*, **12**, 241–248 (1993).
12. Wassell, J. T. and Kulczycki, G. W. 'Frailty analysis of repeated injuries', *Proceedings of the Epidemiology Section of the American Statistical Association*, 130–134 (1995).
13. Klein, J. P. 'Semiparametric estimation of random effects using the cox model based on the EM algorithm', *Biometrics*, **48**, 795–806 (1992).
14. Wang, S. T., Klein, J. P. and Moeschberger, M. L. 'Semi-parametric estimation of covariate effects using the positive stable frailty model', *Applied Stochastic Models and Data Analysis*, **11**, 121–133 (1995).

15. Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. 'A counting process approach to maximum likelihood estimation in frailty models', *Scandinavian Journal of Statistics*, **19**, 25–43 (1992).
16. Pickles, A. and Crouchley, R. 'A comparison of frailty models for multivariate survival data', *Statistics in Medicine*, **14**, 1447–1461 (1995).
17. Clayton, D. G. 'A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic heart disease', *Biometrika*, **65**, 141–151 (1978).
18. Hougaard, P. 'A class of multivariate failure time distributions', *Biometrika*, **73**, 671–678 (1986).
19. Fleming, T. R. and Harrington, D. P. *Counting Processes and Survival Analysis*, Wiley, New York, 1991.
20. Landen, D. D. 'A case-control study of electrical injury among a cohort of line mechanics', *American Journal of Industrial Medicine*, 1999, in press.
21. *S-PLUS* 4 *Guide to Statistics*. Data Analysis Products Division, Mathsoft, Seattle, WA, 1997.
22. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, London, 1993.
23. Shao, J. and Tu, D. *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995.
24. Mackenzie, G. 'A proportional hazard model for accident data', *Journal of the Royal Statistical Society*, *Series A*, **149**, 366–375 (1986).
25. Shih, J. H. and Louis, T. A. 'Assessing gamma frailty models for clustered failure time data' *in* Jewell, N. P., Kimber, A. C., Lee, M-L. T. and Whitmore, G. A. (eds), *Lifetime Data*: *Models in Reliability and Survival Analysis*, Kluwer Academic Publishers, Dordrecht, 1996, pp. 371–379.

STATISTICS IN MEDICINE
Statist. Med. **18**, 3373–3375 (1999)

CLOSING REMARKS†

This spot on the symposium programme traditionally is reserved for Dr. Ted Colton, Professor of Epidemiology and Biostatistics at Boston University's School of Public Health and founding editor of *Statistics in Medicine*. We should acknowledge the role that Dr. Colton has played in these symposia since their inauguration in 1988. It was through his vision that the first of these symposia was held. The proceedings from this conference will mark the 10th anniversary of the collaboration of CDC and *Statistics in Medicine*, published by John Wiley & Sons. His support, encouragement, and constant urging to quality and timely publication have greatly affected the creation and dissemination of the results of these symposia.

Secondly, I thank Ken Falter and his Deputy Chair, Don Betts, for a most commendable symposium. They, along with their excellent staff, have produced a stimulating and valuable conference. Donald and Ken, I welcome you now to the distinguished society of past chairs of CDC/ATSDR statistical symposia.

To summarize the conference, I used an exploratory data analytic approach. A pattern emerged, and in honour of Dr. Colton, who could not be with us for the first time ever in this series, we call these the 'C' concepts. The first concept that I saw emerging from the talks and the posters is *completeness*. The first symposium in 1988 was a symposium on statistics and surveillance. It was held at CDC and approximately 100 people came together and presented about 25 papers. The topic of surveillance was chosen since it was and remains a core public health function. The purpose of the first symposium was to bring to one of the core public health functions the science of statistics and begin to illustrate how statistics can bring an important piece of evidence to that activity. Still in this conference we see statisticians dealing with issues relating to surveillance. We heard about analysis of aberrations, quality control charts from industry, cluster analyses, and Bayesian methods of surveillance analysis.

We have learned much since that first symposium. In 1997, we have approximately three times the number of attendees from academia, industry and international countries. We have representation from 28 states and nine international countries. As a colleague of mine said recently 'So, 300 of your closest friends are coming to town tomorrow, right?' The second symposium had to do with cluster analyses, and we heard about clusters again at this symposium (for example, Dr. Chen's paper on q-statistics and the last parallel session on geographical smoothing and spatial analysis).

The topic of the third symposium on the evaluation of intervention and prevention strategies was revisited here when we heard about the evaluation of hepatitis B immunization programmes and the evaluation of control measures for a measles outbreak. The fourth symposium on the combination of data from multiple sources was recalled by a meta-analysis of educational and intervention activities in schools. And finally, the last symposium on geographical information systems concerned methods used in the plenary session on spatial statistics.

Published in 1999 by John Wiley & Sons, Ltd.

Another aspect of 'completeness' in this symposium regarded *completeness of data*. We heard about missing values in surveys, and our colleagues from North Carolina told us the effect on policy of misclassification of race/ethnicity and other variables in vital statistics data.

The second concept that emerged in this symposium was one of 'consensus'. Dr. David Sencer gave us an example in the opening plenary session of how consensus is not a safety net for shielding us from criticism when our analysis takes us into controversial public health arenas. Dr. Ed Sondik illustrated effectively that statistical analysis alone is not always sufficient for consensus, and Dr. Lee Annest advised us that consensus and policy should be informed by rigour, completeness of analysis and communication. Dr. Sallie Keller-McNulty illustrated the importance of a team effort in changing the standard operating procedures at Los Alomos National Laboratory.

The third concept that arose from the talks was one of '*context*'. This symposium is a vivid reminder that our work as statisticians and analysts in public health is contextual. You can find examples of randomization as it was used in health care as early as the 16th century, when randomization designs were used to evaluate the practice of blood letting, and the use of fruit as treatment and prevention of scurvy. A very real or personal link exists between statistics and public health and one can see it in the work of the Shattuck brothers of Massachusetts in the 19th century. Lemuel Shattuck was an early proponent of using surveillance data, particularly vital statistics, to assess the health of Massachusetts residents. His brother was one of the founding members of the American Statistical Association.

This week we saw statistics in several contexts. Statistical methods were presented in the areas of: regression; time series analysis; Latin squares; various smoothing algorithms; simulation, which Dr. Price cleverly called 'assigning truth'; power and sample size determination; Bayesian analysis; survival analysis, and neural networks. We borrowed strength, if you will, from engineering, economics, ethics, decision modelling and geography.

Finally, the context of the symposium took us into the public health problems of asthma, school health education, gasoline in the cars that we drive, disposal of hazardous waste in the environment, infectious diseases in populations, diabetes, vaccination programmes, and for breast cancer screening and genetic testing. This is a cogent reminder that the consequences of our analyses are the quality and the quantity of health in human populations.

Next, a concept in the symposium was '*communication*'. Our human populations are keenly influenced by concepts of risk. I learned driving home last night that the safety records of U.S. airlines will soon be available on the Internet; before booking your next flight, you will be able to assess historical data on the safety records of every major U.S. airline. However, even the aviation official interviewed on national public radio admitted 'we will not provide a ranking of the airlines' safety record because the numbers of incidents are so small that any difference between them will not be statistically significant'. And yet I would argue that we may not have an expertise in communicating public health risks and prevention messages to two audiences – the communities of people whose health we would like to influence and the communities of policy makers whose decisions we would like to affect.

Consider last week's *Wall Street Journal* headline, 'Shaky statistics are driving the air bag debate'. The discussion described the Senate hearing about data that showed an increase in the number of air-bag-related deaths, especially among children. It reminded Dr. Karen Kafadar of a statement made many years ago by statistician Lincoln Moses: 'White horses eat more hay than black horses. Of course, there are more white horses than black horses'. The field of exploratory data analysis gives us wonderful tools for communication. Just think of some of the vocabulary:

data depth; jittered scatter plots; one wild distributions; interfences; ladders of transformations; and the Hallowe'en effect.

Finally, the last concept that emerged was one of '*challenges*'. When I talked with Dr. Colton this morning, he asked me to bring his greetings to you and advised me to challenge the publication committee to have a timely publication of the proceedings. In the opening plenary, Dr. Broome, Deputy Director, CDC, challenged us to do scientific evaluations of real public health programmes in human populations. And finally, I've heard several challenges that we should merge statistical activities with other methods of science, such as economic methods, to provide confidence bounds and risk estimates for the evaluation of policy decisions. Twelve years ago, when I made the switch from an academic teaching environment to work at CDC, I was trying to explain to my mother the shift in my career. My mother, the British literature professor, listened very attentively and then said 'now let me get this straight … for 12 years I've been telling my friends that my daughter is a statistician, and now you're telling me that I have to tell them that my daughter is an epidemiologist'. The conference challenges us as statisticians and analysts in public health to do two things: to practice the best quality of science that will withstand not only peer review by our colleagues and the editors of *Statistics in Medicine*, but also the scrutiny of *The Wall Street Journal* and the Environmental Protection Agency. And we must engage in presenting the results of these analyses in a way that even my mother's friends will understand.

Thank you for sharing your impressive talents with us. The quality of the posters, the speakers, the moderators, and the audience discussion convinced me that collectively we can rise to these challenges. Thank you for coming, and with Dr. Colton I look forward to seeing the proceedings in *Statistics in Medicine*.

Donna F. Stroup
Centers for Disease Control and Prevention
1600 Clifton Road NE
Mailstop C-08
Atlanta, GA 30333, U.S.A.

Published in 1999 by John Wiley & Sons, Ltd. *Statist. Med.* **18**, 3373–3375 (1999)