League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance

Harvey Goldstein; David J. Spiegelhalter

Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 159, No. 3. (1996), pp. 385-443.

Stable URL:
http://links.jstor.org/sici?sici=0964-1998%281996%29159%3A3%3C385%3ALTATLS%3E2.0.CO%3B2-5

Journal of the Royal Statistical Society. Series A (Statistics in Society) is currently published by Royal Statistical Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Mon May 14 09:57:04 2007

J. R. Statist. Soc. A (1996)
159, *Part* 3, *pp.* 385–443

League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance

By HARVEY GOLDSTEIN† and DAVID J. SPIEGELHALTER

Institute of Education, London, UK *Medical Research Council Biostatistics Unit, Cambridge, UK*

[*Read before* The Royal Statistical Society *on Wednesday, November 15th, 1995, the President,* Professor A. F. M. Smith, *in the Chair*]

SUMMARY

In the light of an increasing interest in the accountability of public institutions, this paper sets out the statistical issues involved in making quantitative comparisons between institutions in the areas of health and education. We deal in detail with the need to take account of model-based uncertainty in making comparisons. We discuss the need to establish appropriate measures of institutional 'outcomes' and base-line measures and the need to exercise care and sensitivity when interpreting apparent differences. The paper emphasizes that statistical methods exist which can contribute to an understanding of the extent and possible reasons for differences between institutions. It also urges caution by discussing the limitations of such methods.

Keywords: MARKOV CHAIN MONTE CARLO METHODS; MULTILEVEL MODELLING; PHYSICIAN PROFILING; RANKING; RISK SHRINKAGE ESTIMATORS; STRATIFICATION; VALUE ADDED

1. INTRODUCTION

Over the last decade there has been an increasing interest in the development of 'performance indicators' as part of an attempt to introduce accountability into public sector activities such as education, health and social services, where the focus has been on the development of quantitative comparisons between institutions. Education is the area where performance indicators seem to have been subject to the longest period of development and use, but more recently in the UK hospitals (National Health Service Executive, 1995) and local government services (Audit Commission, 1995) have had attention. Smith (1990) discussed the background to this interest and looked at the social, economic and political purposes performed by performance indicators in both the private and the public sectors. In contrast, the present paper focuses on statistical methodology, and, although we shall be offering suggestions about appropriate ways of modelling and interpreting performance indicator data, our principal aim is to open up a discussion of the issues rather than to prescribe specific solutions to what are clearly complex problems.

In its widest sense a performance indicator is a summary statistical measurement on an institution or system which is intended to be related to the 'quality' of its functioning. Such measures may concern different aspects of the system and reflect different objectives: 'input' indicators such as the pupil/teacher or the staff/bed ratio

†*Address for correspondence*: Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK.
E-mail: hgoldstn@ioe.ac.uk

© 1996 Royal Statistical Society 0035–9238/96/159385

are often used to estimate the resources available to institutions, 'process' measures such as average teaching time per pupil or proportion of day case surgery may reflect organizational structure, whereas 'outcome' measures such as school examination results or hospital operative mortality have been used to judge institutional 'effectiveness'. Although much of our discussion is relevant to input and process measures, we shall concentrate on outcome indicators in the areas of education and health, since it is these which have come to assume a considerable social importance and a central role in political debates about institutional accountability.

There is a very important debate over the best choice of indicator measures and their validity as measures of effectiveness: within both education and health we shall cast doubts on whether variability in outcomes, even after adjusting for external factors, does adequately reflect the 'quality' of institutions. Appropriateness of an indicator will involve practical as well as social and political considerations and we make no attempt at a systematic review of the large relevant substantive literature. By concentrating on statistical issues we do not intend to belittle the importance of such concerns, but we believe that the statistical procedures we discuss are generally applicable whatever measures are chosen and for whatever purpose. We also believe that all potential users of performance indicators should have at least a basic awareness of these issues, whether these users are, for example, hospital administrators or parents of school-children.

In emphasizing a statistical perspective that is common to any subject-matter area, we shall argue for the proper contextualization of outcome indicators by taking account of institutional circumstances and the appropriate specification of a statistical model. We shall also stress that there are quantifiable uncertainties which place inherent limitations on the precisions with which institutions can be compared. Although many technicalities are relevant to our discussion, we shall try to present it in a relatively informal fashion, with suitable references to the relevant literature.

The structure of the paper is as follows. We first provide a brief overview of the use of performance indicators in education and health, and in Section 2 we describe a basic framework for some issues that are common to all attempts to compare institutions by using outcome data. The discussion of statistical modelling is then expanded into a section which contains the limited technical aspects (and hence may be skipped at the reader's discretion). Sections 4 and 5 deal with education and health respectively by using practical examples. Although there are inevitable differences in background and approaches of the authors, the commonality of view is dominant. The final section attempts to bring the discussion back to common themes, and summarizes our opinions on the possible future role of 'league tables'.

Before we introduce a discussion of the detailed issues and a presentation of the research evidence a remark about terminology is necessary. In judging whether an institution has enhanced the welfare, health or performance of its members between their entry and exit, the term 'value added' has come to be used widely. This term is borrowed from economics, but it is difficult to justify in areas such as education or health. It is rarely the case that inputs and outputs are measured in the same units, as would be the case for example when dealing with monetary costs. If we are judging institution aggregate examination scores for example, the intake variable will typically be quite a different measurement of achievement, ability etc. Likewise, for measurements such as surgical success rates, we would wish to contextualize these by

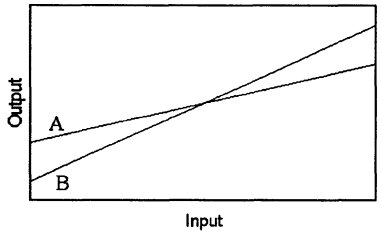


Fig. 1. Simple regression relationships of output on input for two hypothetical institutions

using measures such as the severity of the presenting condition. It follows that there is no absolute scale of measurement whereby we can measure how many units an institution has 'added' to its pupils or patients. Rather, our comparisons will be relative where institutions will be judged against each other, conditional on prior or base-line measurements, and so the indicators that we shall refer to in this paper are 'comparative'. In some cases, e.g. immunizations, it may be possible to achieve agreement on 'absolute' standards of performance. Although similar statistical issues will arise in those cases, we shall not be concerned with them. The typical, if somewhat oversimplified, case is exemplified in Fig. 1 which shows a comparison between two institutions in terms of predicted mean output for a given input or base-line measure. Institution A has a higher expected achievement than B for individuals with low base-line scores and vice versa for those with high base-line scores. The issue therefore is the comparative issue of whether one institution has a higher expectation than another, for given values of adjustment factors, remembering that there may be several possible adjustment factors. Rather than value added, therefore, we prefer the term 'adjusted comparison'.

1.1. *Performance Indicators in Education*

The Organisation for Economic Co-operation and Development (OECD) (1992) has been active in developing sets of educational performance indicators for national systems which include measures of students' achievement and which they see as complementing similar indicators at the level of institutions within national education systems. This activity is supported by the member states of the OECD and reflects a very general concern with ways in which institutions and systems can be judged. In justifying this concern the OECD refers, for example, to local education authorities who 'require data on the patterns of educational outcomes across schools . . . for potential use in decision making'. The OECD also identifies a shift from the use of input indicators such as expenditure to a concern with outputs such as student achievement. Interestingly, the report is little concerned with process indicators such as curriculum organization or teaching styles and we shall have more to say about such variables later in both health and education.

Although the focus of the present paper is on the use of outcome indicators for comparing institutions, or indeed whole educational systems, essentially the same issues arise whichever level of aggregation is of concern. Thus, the OECD report appears to assume that comparisons of students' achievements across countries, unadjusted for context, allow inferences about the relative performances of educational systems. Such an assumption has pervaded almost all existing discussions of international comparative data (Goldstein, 1995). Until recently, in the UK, this

was also the implicit assumption behind the publication of examination results in the form of institutional rankings or 'league tables'. Yet at the intranational level the debate has now shifted markedly from this simplistic assumption towards a recognition that institutional and subsystem comparisons must be contextualized, principally by making adjustments for student status and achievements on entry to the education system or particular phases of it (Sanders and Horn, 1994; Fitz-Gibbon, 1992; Department for Education, 1995a).

In the UK the debate has centred on the notion of adjusted (value-added) comparisons between schools and other educational institutions. In a briefing paper, the Department for Education (1995a) proposed that a base-line of prior attainment should form the adjustment measure for any particular phase of schooling and pointed out that this will be a better reflection of a school's contribution to the performance of its pupils. The assumption is that, if suitable measurements, data collection strategies, etc. can be devised then it will become possible for 'the performance of schools and colleges to be compared consistently across the country'. We argue below, however, that such an aim, although worthy and an improvement on unadjusted 'raw' league tables, is generally unrealizable: adjusted league tables inherit many of the deficiencies of unadjusted ones and an appreciation of well-established statistical principles of uncertainty measurement needs to inform public debate.

1.2. *Performance Indicators in Health*

In contrast with the educational domain, the term 'institution' needs to be given a broad definition to cover the application of indicators to health outcomes. Three levels can be distinguished, broadly defined in terms of health authorities, hospitals and clinicians: the examples in Section 5 have been deliberately chosen to illustrate non-hospital comparisons.

1.2.1. *Health authorities (the purchasers)*

Indicators for avoidable mortality in area health authorities (Charlton *et al.*, 1983) and the public health targets established in the 'Health of the nation' programme (National Health Service Management Executive, 1992) have been developed into a set of population outcome indicators for the National Health Service (NHS) calculated at the purchasing authority level (McColl and Gulliford, 1993). These are now distributed in printed and disc form as the public health common data set (Department of Health, 1994) in which regions in England are ranked for each indicator, and there is growing emphasis on appropriate means of assessing whether local areas are progressing towards, or achieving, national or local targets (National Health Service Management Executive, 1993). In Scotland, the linked medical record system has permitted greater progress, including published tables on a variety of outcome measures both for health authorities and for trusts (Scottish Office, 1994). Methodology employed in all this work will be discussed in Section 5.2.1.

The current practice of using activity and finance data to calculate an 'efficiency index' for each purchasing authority lies outside our immediate concentration on outcome measures, although it is likely that this procedure would also benefit from an additional acknowledgement of uncertainty.

1.2.2. *Hospital trusts (the providers)*

Before the reorganization of the NHS, most hospital outcome data were aggregated to the district level before forming part of a long list of indicators each of whose rank across districts could be graphically displayed (Yates and Davidge, 1984; Lowry, 1988). Currently there is public dissemination of such process measures as waiting times and adherence to appointment times (National Health Service Executive, 1995). There has been considerable public debate surrounding each such publication: institutions tend to be immediately ranked by the media and the apparently 'best' and 'worst' become subject to close scrutiny, accompanied by criticism from clinicians and statisticians of the naïve interpretation of the results as reflecting the 'quality' of care—Richard Rawlins is reported as saying 'We should insist on correct political arithmetic, not arithmetic political correctness' (British Medical Journal, 1995). Past comparisons between the outcomes achieved by different hospitals in the UK have generally been strictly anonymized, such as confidential audits of cardiac surgical mortality and perinatal deaths (Leyland *et al.*, 1991), although public dissemination of hospital-specific outcome measures appears inevitable and, as mentioned earlier, is already occurring in Scotland.

In the USA programmes concerning, for example, mortality of Medicare patients (Jencks *et al.*, 1988) and cardiac surgical outcomes (Hannan *et al.*, 1994) have developed amid criticism of inadequate adjustment for the type and severity of illness of patients, the poor quality of the data and the possibility of systematic manipulation by institutions. However, recent discussion on these programmes argues that they are maturing into reasonable tools for quality control and audit in that, for example, Medicare studies are being based on detailed patient characteristics rather than simple adjustment for routinely available age, sex and comorbidity factors. For an extensive discussion of issues surrounding comparisons of hospitals, see the recent special issue Annals of Thoracic Surgery (1994).

1.2.3. *Individual clinicians*

There is strong resistance to explicit identification of individuals and their associated outcome statistics, although occasional anonymous comparisons have been published (e.g. McArdle and Hole (1991)). However, the New York State cardiac mortality programme features named surgeons, and this is discussed further in Section 5.2.2.

2. COMMON ISSUES IN PERFORMANCE COMPARISONS

The following framework sets the structure for the succeeding discussion within the contexts of education and health, and separates common issues into those concerned with the data collected, technical aspects of statistical modelling and presentation, and finally the interpretation and possible institutional effect of such comparisons.

2.1. *Data*

No amount of fancy statistical footwork will overcome basic inadequacies in either the *appropriateness* or the *integrity* of the data collected. For example, appropriate and relevant outcome measures are controversial, especially within the health

context, as well as the selection and measurement of confounding factors for which the outcomes may need to be adjusted. Integrity of data covers not only basic quality issues of completeness and correctness but also the possibility of deliberate manipulation.

2.2. *Statistical Analysis and Presentation*

We shall pay particular attention to the specification of an appropriate statistical *model*, the crucial importance of *uncertainty* in the presentation of all results, techniques for *adjustment* of outcomes for confounding factors and finally the extent to which any reliance may be placed on explicit *rankings*. The technical aspects of these are dealt with in the following section.

2.3. *Interpretation and Impact*

The comparisons discussed in this paper are of great public interest, and this is clearly an area where careful attention to limitations is both vital and likely to be ignored. Whether adjusted outcomes are in any way valid measures of institutional 'quality' is one issue, while analysts should also be aware of the potential effect of the results in terms of future behavioural changes by institutions and individuals seeking to improve their subsequent 'ranking'.

3. STATISTICAL MODELLING, ANALYSIS AND PRESENTATION

3.1. *Models*

We shall discuss the use of outcome indicators within the framework of multilevel model fitting. The data structures that we are concerned with are hierarchical in nature, patients being nested within hospitals and students within schools. In practice, real data structures may often be more complex, involving spatial and other factors. For example, patients are not only nested within hospitals, but the latter are 'crossed' with localities and general practitioners. If general practitioners are influential then they should be incorporated into the statistical model if trying to estimate an effect associated with institutional performance. In education, there is evidence (Goldstein and Sammons, 1996) that examination results at the end of secondary schooling are influenced by the primary or elementary school attended, so that students need to be cross-classified by secondary and primary school. Likewise, interest may focus on institutional trends over time, such as monitoring progress towards the health of the nation targets, and this adds further modelling complexity. For simplicity of exposition we shall not deal with these cases: technically there are straightforward extensions to the existing procedures for handling purely hierarchical data (Goldstein, 1995).

Our use of multilevel models reflects our default assumption that having made suitable adjustments we expect institutions broadly to be similar. Statistically this means that higher level units can be regarded as drawn from a population of units or, more technically, to be 'exchangeable' (Bernardo and Smith, 1994). Interest centres both on the between-unit variation and on posterior or predicted estimates of unit effects. The latter estimates are the familiar 'shrunken' estimates which have the useful property of moving higher level unit estimates towards the population mean value and increasing precision and accuracy of prediction (see, for example, Morris

(1983)). Bayesian or maximum (quasi-)likelihood estimates are readily obtained and in the following data analyses we have used Gibbs sampling for the former (Gelfand and Smith, 1990) and iterative generalized least squares (Rasbash and Woodhouse, 1995; Goldstein, 1995) for the latter. What we have to say applies whether responses are continuous normally distributed data, counts, proportions or, for example, survival times, and statistical preferences between Bayesian, likelihood and quasi-likelihood methods are usually more of philosophical than practical importance.

For simplicity consider a basic model for a single-year cohort of students nested within schools and on whom we have an examination score as response (Y). We can write a two-level variance components model

$$\left.\begin{aligned} y_{ij} &= \beta_0 + u_j + e_{ij}, \\ \mathrm{var}(u_j) &= \sigma_u^2, \\ \mathrm{var}(e_{ij}) &= \sigma_e^2 \end{aligned}\right\} \tag{1}$$

where y_{ij} is the examination score for the ith student in the jth school, u_j is the residual or 'effect' for the jth school and e_{ij} the residual for the ith student in the jth school. The residuals are assumed mutually independent with zero means. Given student level data this model can be fitted as indicated above and in particular will yield posterior estimates $\hat{u}_j$ and $\mathrm{var}(\hat{u}_j)$ or alternatively $\mathrm{rank}(\hat{u}_j)$ and $\mathrm{var}\{\mathrm{rank}(\hat{u}_j)\}$, which in turn can be used for comparisons between institutions. We shall discuss below exactly how these can be used. In the health applications the lowest level units are patients and the higher level units are physicians or hospitals. The extension of our methods to three-level and higher level models, and also to models with cross-classifications of units, is straightforward.

It is important to fit a model in which institutional differences are modelled explicitly. Failure to do this will result in biased inferences arising from the lack of independence induced by the multilevel structure. It may also result in serious model biases, especially where the underlying structure is more complex than a variance components model, e.g. in the common case where there are random coefficients at the level of the hospital or school.

3.2. *Uncertainty and Institutional Rankings*

We shall repeatedly emphasize the need for interval estimation in which the uncertainty associated with estimates or ranks is explicitly displayed. Regardless of the care with which the statistical analysis is carried out, it is inevitable that the resulting point estimates will lead to institutional ranking or league tables. However, although such ranks are particularly sensitive to sampling variability, there has been no straightforward way to place interval estimates around those ranks. Fortunately, modern computing technology allows Monte Carlo estimates to be obtained by simulating plausible estimates and hence deriving a large sample of possible rankings which can be summarized by, say, 95% intervals: maximum (quasi-)likelihood models lend themselves to bootstrapping samples, whereas Markov chain Monte Carlo techniques (Besag *et al.*, 1995; Spiegelhalter *et al.*, 1995) easily accommodate the ranking of the set of parameter realizations at each iteration and hence the reporting of point and interval estimates of the ranks alongside the parameter estimates.

We have used two procedures for deriving intervals. The first procedure, illustrated in the examples of Section 4, was proposed by Goldstein and Healy (1995) and provides, for each institution, an interval centred on the mean, and two institutions are judged to be statistically significantly different at a preassigned level for their means if and only if their respective intervals do not overlap. The procedure has the property that the average type 1 error, over all possible equally likely pairwise comparisons, is at the specified level. The procedure can be extended to allow for multiple comparisons among, say, triplets of institutions, and in this case the interval estimates will generally become wider. Its use will be appropriate when each member of a class of users is concerned only with the comparison of two (or more) particular institutions. This would be the case, for example, if all parents were concerned to choose between two or three locally accessible secondary education institutions for their children.

The second procedure, presented in Section 5, is to apply conventional, say 95%, intervals around the mean for each institution. For any particular institution this locates it within the overall population distribution. For intervals constructed on the response variate scale the population distribution can be estimated directly from the (shrunken) distribution of the posterior residual estimates. An alternative, which we have adopted in this paper, is to display the ranked residuals together with intervals on these ranks. This has the advantage of being more readily understood by non-specialists, although in general we will obtain relatively wider intervals. We note that it would be possible to adapt the first procedure to be displayed in terms of rankings, but we have not done this because interest there centres on comparisons between specific institutions.

3.3. *Adjustment*

The need to adjust for initial status has been strongly argued within both education and health, and this can be accommodated in two ways. First, subject-specific covariates may be included in the generalized linear models described in Section 3.1 — the appropriateness of also including institution-specific covariates will be discussed separately for each context. The second approach exploits an existing adjustment procedure to derive an expected aggregate outcome for each institution based on the characteristics of its intake, and then a residual is based on the contrast between observed and expected outcomes. This latter 'risk stratification' approach is widely adopted in medical studies since the adjustment system may be published and applied in prospective assessment in new institutions and avoids continual reanalysis of an entire data set.

4. EDUCATION

4.1. *Data*

The most commonly used measurements for institutional comparisons in education are test scores or examination results. In the UK, the latter have been the principal focus, but it is intended that the results of national curriculum assessments will be used in the future. In addition, there is interest in other outcomes such as student attitudes and attendance. To the extent that these also are influenced by extra-institutional factors such as social background, prior behaviour and achievement, then they need to be contextualized suitably.

A common obstacle to carrying out appropriate adjustments when modelling examination results is the lack of suitable prior achievement measures. Our first educational example is of such an unadjusted case, while the second is for A-level General Certificate of Education (GCE) examinations where earlier General Certificate of Secondary Education (GCSE) examination results are available.

4.2. *Statistical Analysis and Presentation*

4.2.1. *Uncertainty estimation*

To illustrate the problem we shall consider a set of simple, unadjusted A-level GCE results from one local education authority for the years 1993 and 1994. These are extracted from tables published by the Department for Education annually (Department for Education, 1994, 1995b). We have chosen to use data from one local education authority but the points that we wish to make will apply more generally.

For the present purposes we have excluded schools which have a selective intake and included only maintained schools. We use the standard A- and AS-level point scoring system whereby an A-grade is allocated a score of 10, a B-grade a score of 8 etc. with half these values for corresponding grades for the AS-examination. We shall refer to these scores simply as A-level scores. Each student then has a total score formed by summing the separate subject scores. The number of students per school per year ranges from 8 to 81 with a mean of 44. We wish to display the mean scores for all schools and at the same time to display a measure of uncertainty.

We recognize the limitations of an aggregate A-level score across all subjects. Institutions certainly differ in terms of the relative 'effectiveness' of different subject departments and this is recognized in the A-level information system (Fitz-Gibbon, 1992) which we shall return to in the discussion. One difficulty, however, is that for many departments the numbers of students following an A-level course will be small, resulting in very wide uncertainty intervals.

In the present case we do not have data for individual departments nor are the student level data published, but we do have the total number of students entered for examinations in each school. We fit the following model for the mean total A-level score, based on aggregating model (1):

$$\left.\begin{aligned} y_j &= \beta_0 + u_j + e_{.j}, \\ e_{.j} &= \sum_{i=1}^{n_j} e_{ij}/n_j, \\ \operatorname{var}(e_{.j}) &= \sigma_e^2/n_j. \end{aligned}\right\} \tag{2}$$

Since we know the n_j we can estimate the student level variance from the present data (95.2) and this estimate is not too different from that obtained (78.0) by Goldstein and Thomas (1996) using student and school level data in a large scale analysis of A-level results in 1993.

Since we have 2 years of data we can study both the average over 2 years and the difference; the latter is of interest to see whether schools can be distinguished in terms of changes over time. We therefore fit the following two-level model:

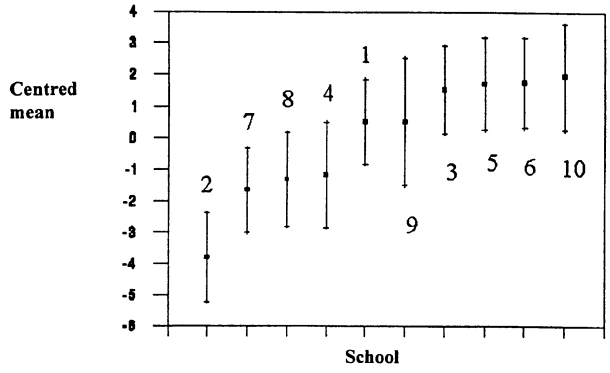


Fig. 2. School intercept residual estimates and 95% overlap intervals

$$\left.\begin{aligned} y_{hij} &= \beta_0 + \beta_1 z_{hj} + u_{0j} + u_{1j} z_{hj} + e_{hij}, \\ z_{hj} &= \begin{cases} 1 & \text{if } h = 2, \\ 0 & \text{if } h = 1 \end{cases} \end{aligned}\right\} \tag{3}$$

for the ith student in the jth (1, 2) school in the hth (1, 2) year. This model is then aggregated to the school level as before. Thus u_{0j} and u_{1j} respectively are the jth school (intercept) effect and difference between year 2 and year 1 with associated variances and covariance. Figs 2 and 3 show the estimates, which are approximately independent, ordered for the set of schools: Fig. 2 shows estimates of the school averages and overlap intervals and Fig. 3 shows estimates of the year 2 – year 1 differences for each school with overlap intervals. The school numbers are displayed on each figure.

In terms of pairwise comparisons, whereas school 2 could be distinguished from each of the highest six schools, and school 7 from the highest four, the others cannot be separated among themselves. When we look at the year differences we see that none of the schools can be separated in this way. The latter result is also found to hold for trend estimates of up to 5 years (Gray *et al.*, 1995).

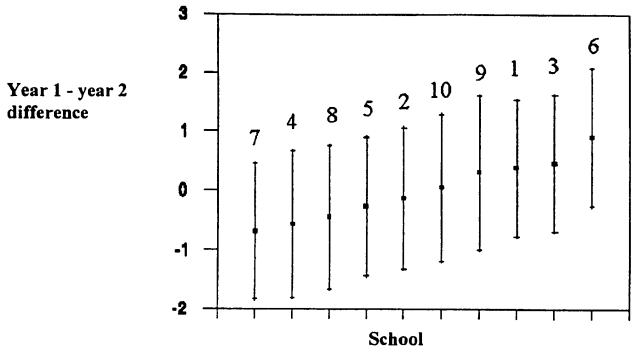


Fig. 3. Year difference residual estimates for each school and 95% overlap intervals

TABLE 1
Parameter estimates from fitting model (3)†

Parameter	*Estimate*	*Intercept*	*Year*
Fixed			
Intercept	15.8		
Year	−1.4 (0.7)		
Random between schools (variances and covariance)			
Intercept		4.06 (3.2)	
Year		−0.21 (2.5)	1.69 (4.3)

†Standard errors are given in parentheses.

Table 1 gives the parameter estimates from fitting model (3). We notice the imprecision of these estimates based on a small number of schools. We see therefore that even before carrying out adjustments, using existing published data on raw results, there are strict limitations to the comparisons which it is statistically legitimate to make. It is also our view that presentations such as in Figs 2 and 3 which compare institutions with one another are the appropriate presentations to make. It is sometimes suggested that each institution should be compared with the sample mean and judged by whether its separate (conventional) uncertainty interval overlaps this mean. We can see little justification for this, since the mean has no special status, and we can of course have two institutions, both of which have, say, 95% intervals overlapping the mean but with an uncertainty interval for the difference which does not overlap 0. Where the principal focus is on comparisons of particular institutions it is appropriate to present these directly. We also note, as pointed out earlier, that where more than two institutions are compared diagrams such as Figs 2 and 3 present a conservative picture as they are designed only for pairwise comparisons.

4.2.2. *Adjustment procedures*

There is now a large body of evidence which demonstrates that the single most important predictor of subsequent achievement in school is obtained by using measures of intake. For example, in their classic paper 10 years ago, Aitkin and Longford (1986) obtained a correlation of about 0.7 between examination results at 16 years of age and verbal reasoning scores at 11 years at the start of secondary school. Because of the selective nature of most educational systems it will almost always be necessary to adjust for such intake variables. Research has also shown (see for example Goldstein and Thomas (1996)) that the adjustment relationship may be complex and in particular that there are interactions with other factors such as gender or social background. Furthermore, when multilevel models are fitted allowing for variation at both the institutional and the student level, this variation is also complex. Thus, the between-school variation in A-level scores is higher among the low scoring students at GCSE, and also for girls, and the institutional effect varies by both gender and GCSE score.

Raudenbush and Willms (1995) distinguished two kinds of adjusted institutional comparisons. They label as 'type A' comparisons those which are primarily of

interest to students and others concerned with choosing between institutions. For such individuals they wish to ascertain the expected output achievement conditionally on their own characteristics, such as their input achievement, social background, gender etc. They will also be interested in whether there are interactions between their own characteristics and those of the other students likely to attend any institution. Thus, for example, there is some evidence (Steedman, 1980) that, at certain intake achievement levels, attendance at a secondary school where the average intake score is higher than the student's leads to a raised output score compared with attendance at a school where the average is lower.

Type B effects are those where, in addition to the type A effects, we are interested in measuring those characteristics of institutions and their members which further explain institutional differences. Thus, curriculum organization or streaming policy may explain some of the variation in outcome and this may help us to construct causal explanations. Strictly speaking, in choosing an institution, the potential student need not known *why* institutions differ. Nevertheless, the reasons for differences are of interest generally for accountability. For example, suppose that schools which stream strongly enhance the achievements of students with high intake achievement but depress the achievements of those with low intake scores. Explanations for such effects will be of interest to those such as local education authorities who are responsible for promoting the progress of all students. This example also raises the interesting issue of feed-back effects, namely that public knowledge of this kind which then is acted on may change the composition of institutions and hence directly affect their operation, so that the relationships formerly observed no longer hold. We shall return to this issue.

In the absence of good understandings about type B effects, the distinction between type A and type B effects is of little practical significance, although the study of type B effects remains an important research topic.

To illustrate the effect of adjusting for inputs we use results from Goldstein and Thomas (1996) based on the analysis of A-level and GCSE results for 325 schools and colleges and 21 654 students. Fig. 4 plots the centred mean A-level score for each school or college against the posterior adjusted estimate for students who score between the lower and upper quartiles of the GCSE distribution. The correlation for

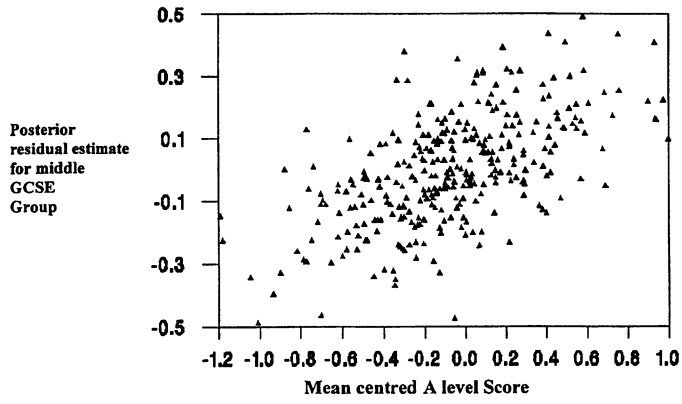


Fig. 4. Residual estimates for the middle GCSE group by mean A-level score (standardized scores)

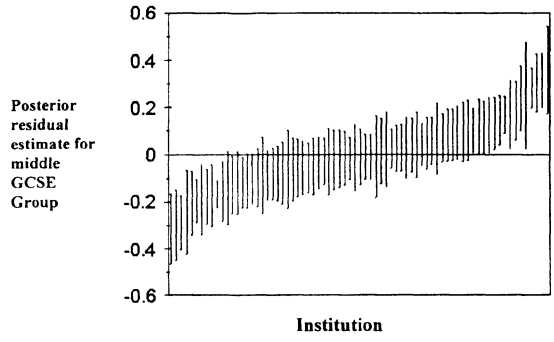


Fig. 5. Pairwise 95% overlap intervals based on adjusted residuals for the middle GCSE group (standardized scores)

Fig. 4 is 0.59, illustrating clearly that there are institutions with relatively high adjusted values who have low raw means and vice versa. If we accept that some form of adjustment is a necessary (although not necessarily sufficient) prerequisite for fair comparisons then failure to use this will, *prima facie*, result in some highly inaccurate inferences.

Fig. 5 shows a random sample of 75 overlap intervals for these institutions based on the posterior residual estimates for the same group of students after adjusting for the GCSE score. From Fig. 5 we can estimate that about two-thirds of all possible comparisons do not allow separation. Thus, even with input adjustment, the use of rankings to judge differences between institutions will have a limited utility. A ranking such as that in Fig. 5 may allow us to isolate some institutions, at the extremes, as possible candidates for further study. In other words we can use such rankings as screening instruments, but not as definitive judgments on individual institutions. The analysis of Gray *et al.* (1995), which uses GCSE results as outcomes with adjustment for intake achievement at 11 years, confirms a similar picture for both a 5-year average and a 5-year trend.

4.3. *Interpretation and Impact*

We have already demonstrated that, with current data, even after adjustment, finely graded comparisons between institutions are impossible. Nevertheless, it is possible that in certain circumstances we might be able to achieve better adjustments and hence more accurate comparisons. Although this cannot be ruled out, and is certainly a legitimate area for further research, there seem to be some inherent limitations to such a process. The principal limitation is that of sample size. The uncertainty intervals for the A-level scores are based on the size of the cohort in any one year. Of necessity this will be small in small institutions. Moreover, if we produce estimates for individual subjects at A-level, the cohort taking a subject in an institution may be very small indeed, leading to wide intervals. It is worth emphasizing that we are regarding the set of students taking an examination as if they were a sample from a superpopulation since we wish to make inferences about the general 'effects' of institutions for *any* group of students in the future.

Any inferences about institutional differences are no better than the data which are

used and the models fitted to them. There are several areas where it is fairly clear that current models are inaccurate. Firstly, measures of input used to date have been opportunistic in that they happen to have been available. The use of verbal reasoning or reading achievement to adjust for overall GCSE score is debatable and it should be possible to improve on the use of a total GCSE score to adjust A-level scores. Secondly, as has already been mentioned, recent research (Goldstein and Sammons, 1996) has shown that when GCSE is the output it is inadequate to adjust solely for achievement at 11 years of age and that information about the junior school attended is important and explains a considerable amount of the between-secondary-school variation. There is also the problem of accounting for students who change schools, who may have particular characteristics, and there is almost no research into this problem. Finally, there is as yet no serious attempt to make adjustments for measurement errors in either the response or the predictor variables in these models. If this is done we would expect possibly substantial changes in parameter estimates and comparisons between institutions (Woodhouse *et al.*, 1996).

Any comparison between institutions on the basis of A-level results is inevitably out of date. If adjustments are made by using GCSE results then a comparison can apply, at the earliest, to a cohort about to enter an institution 3 years after the cohort for whom the results are available. This is the case whether the comparisons are adjusted or unadjusted. For GCSE results this lag is 6 years. If results are aggregated over, say, 3 years, then for A-levels the lag is between 3 and 5 years. The same problem applies to the use of trend data. Institutions can change rapidly, and one way in which this can occur is if potential students actually take decisions on the basis of previous results.

If students decide to choose A-level institutions (where a choice is realistic) on the basis of previous adjusted comparisons then those institutions with apparently 'better' results will have greater competition to enter and will therefore be able to exercise selection. For many this will change the characteristics of their students and the characteristics of the students of 'competing' institutions. If there are interactions between student characteristics and institutional policies then future adjusted comparisons will also change. Given current knowledge the extent and direction of such changes will be difficult to predict, and research into such effects would be important to carry out. For the students, however, the uncertainty raised by these issues is important and may cause them to give only a small amount of weight to comparisons of institutional performance when making choices.

5. HEALTH

5.1. *Data*

Although not the main emphasis of this paper, it is important to note that the vital issues of the appropriateness and quality of data have been discussed at length in the context of assessing an institution's contribution to health outcomes in the NHS, keeping in mind our broad definition of 'institution' as covering both purchasers and providers; see, for example, McColl and Gulliford (1993) and Orchard (1994). Problems include the following.

(a) *Relevance of the population being studied*: in-hospital mortality following admission for myocardial infarction (Scottish Office, 1994) may depend more

on the mix of patients reaching hospital in the first place, rather than the quality of care given once admitted.

(b) *Precise definition of the population under study*: in comparing, for example, 30-day mortality after emergency admission for stroke (Scottish Office, 1994), rates may depend both on the definition of stroke in terms of the international classification of disease codes (ICD9) and the consistency of such ICD9 coding across institutions.

(c) *The definition of the outcome*: 30-day mortality is obtainable in the Scottish analysis because of their record linkage scheme, whereas in-hospital mortality is used in the USA where such routine linked follow-up does not exist. In-hospital mortality in particular may be prone to bias and manipulation, as in the reported tendency of Californian hospitals to discharge patients early whose subsequent 30-day deaths do not count as negative outcomes (McKee and Hunter, 1994).

(d) *Selection and definition of confounder variables*: measures of severity of illness at admission to hospital have been criticized for not fully taking into account known discrepancies in outcomes associated with social background and other factors.

(e) *Quality and completeness of data*: McKee and Hunter (1995) identified problems with routine sources of adjustment data, whereas demanding specially collected severity data brings its own quality control difficulties. Again, this has been extensively discussed within the US cardiac community (Annals of Thoracic Surgery, 1994).

(f) *Deliberate manipulation of data*: this is covered in Section 5.3.

5.2. *Statistical Analysis and Presentation*

5.2.1. *Models*

The use of multilevel or hierarchical models has been pioneered in two areas closely related to institutional comparisons. The first concerns the utilization of different medical interventions, where McPherson *et al.* (1982) provided one of the earliest applications of such 'empirical Bayes' analyses when comparing the use of common surgical procedures in districts in Norway, England and the USA. More recently, Gatsonis *et al.* (1993, 1995) employed random coefficient logistic regression models in comparing rates of coronary angiography between US states, in which the influence of the patient level factors is not assumed constant over all states. The second area concerns the mapping of disease incidence, in which Clayton and Kaldor (1987) again applied empirical Bayes techniques to obtain more accurate estimates of cancer incidence in small areas.

Examples of the use of multilevel models in institutional comparisons include Thomas *et al.* (1994), in their analysis of mortality in Medicare patients, and Leyland and Boddy (1995) when comparing lengths of stay in Scottish hospitals. Closest to our approach is Normand *et al.* (1995), who included both patient and institutional factors within a hierarchical logistic regression model for mortality of Medicare patients, and whose use of Markov chain Monte Carlo methods allows the calculation of any summary measure thought appropriate: for example, as an indication of a possible outlier they calculate the probability that an institution's adjusted rate is more than 50% greater than the median over all institutions.

5.2.2. *Uncertainty and ranking*

Medical performance indicators have shown a traditional emphasis on ranking (Yates and Davidge, 1984; Lowry, 1988), and current publications of process measures attribute 0–5 'stars' to trusts with no comment on uncertainty (National Health Service Executive, 1995). Publications associated with the public health common data set show no interval estimates and all their graphics consistently show regional health authorities in rank order for all outcomes (Department of Health, 1994): data published on disc do provide confidence intervals. Scottish data (Scottish Office, 1994) present confidence intervals in all graphics, although smaller institutions are not shown in graphs. We note that one of the important consequences of using multilevel models should be that a suitable adjustment is made for size so that large and small units can be simultaneously presented.

To focus on issues of estimation and ranking rather than adjustment, and also to show an application using Poisson count data, we shall illustrate the presentation of unadjusted outcomes by using data from the Scottish outcomes study (Scottish Office, 1994) on teenage (ages 13–15 years) conception rates in the period 1990–92 in areas under different health boards: we note that one of the health of the nation targets is to reduce such rates in England to 4.8 per 1000 by the year 2000 (National Health Service Management Executive, 1992). Fig. 6(a) shows the health boards ordered by observed rates and 95% confidence intervals assuming an independent Poisson model within each board, as well as the consequences of adopting a multilevel model in which a Gaussian population distribution is assumed with locally uniform (but just proper) priors on the population mean and log(population variance) (specifically, the population mean and inverse variance were given normal(0, 100 000) and gamma(0.001, 0.001) priors respectively). This random effects model has the predictable consequences of shrinking the point estimates towards the overall mean, reducing the width of the intervals. These and all other estimates are based on empirical summaries of simulated parameter values obtained from 5000 iterations of the Gibbs sampler; satisfactory convergence had been obtained after discarding an initial 1000 iterations.

Using the BUGS Gibbs sampling software (Gilks *et al.*, 1994; Spiegelhalter *et al.*, 1995) it is straightforward to obtain median estimates and 95% intervals for the ranks and these are displayed in Fig. 6(b): the medians do not always match the observed ranks. The width of the intervals is notable: in fact the firmest conclusions that can be drawn are that the Western Isles is in the lower quarter, Highland and Lanark are in the lower half and four health boards are in the top half. The multilevel model, in spite of making the individual estimates more accurate, has the effect of making the ranks even more uncertain, with a particularly strong influence on the Western Isles.

Such unadjusted comparisons appear of limited value in view of the known social class gradient of this and other outcome measures: we now discuss such an adjustment with regard to operative mortality.

5.2.3. *Adjustment*

There is a long history of the development of adjustment procedures for initial disease severity, with an emphasis on cardiac surgery and intensive care, in both of which a range of competing systems exists (Iezzoni, 1994). Recent applications in

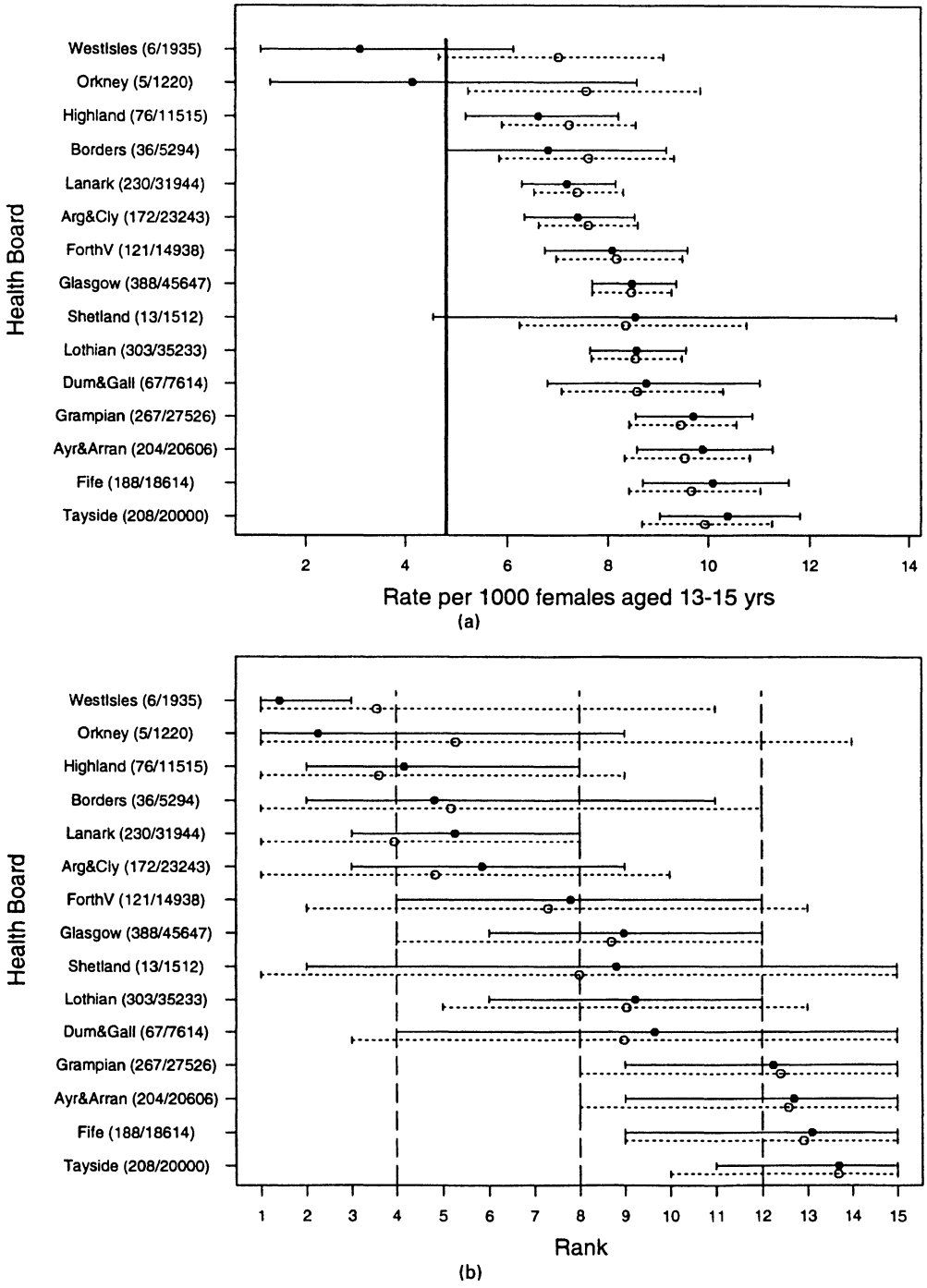


Fig. 6. (a) Estimates and 95% intervals for teenage conception rates assuming independent (———) and exchangeable (·····) Scottish health boards (the English health of the nation target of 4.8 per 1000 is shown); (b) mean and 95% intervals for rank of health boards assuming independent (———) and exchangeable (·····) boards (fixed effects —●, mean; |——|, 95% confidence interval; shrunk estimates —○, mean; |·····|, 95% confidence interval)

which UK institutions have been explicitly (although anonymously) compared include the survival of premature babies by using the CRIB scoring system (de Courcy-Wheeler *et al.*, 1995) and a comparison of survival in intensive care units by using the APACHE II scoring system (Rowan *et al.*, 1993). Risk stratification schemes have been used for adjusting for 'case mix' when measuring change within institutions (see for example Rogers *et al.* (1990)) although here we concentrate on between-institution comparisons.

The New York State Department of Health programme on cardiac artery bypass graft (CABG) surgery seeks to create a cardiac profile system which assesses the performance of hospitals and surgeons over time, independent of the severity of individual patients' preoperative conditions (New York State Department of Health, 1993), and one of its explicit aims is 'providing information to help patients make better decisions about referrals and treatment decisions'. The programme has been recently reviewed by Green and Wintfeld (1995), who described how the publication in the *New York Times* in December 1990 of a league table of hospital CABG-related mortality was closely followed by an appeal by Newsday under the Freedom of Information Act for publication of death-rates according to named clinicians: these were published the following year although only surgeons carrying out more than 200 operations in a single hospital during that period are given by name.

Table 2 shows a sample of the data published in 1993 covering operations for 1990–92: we consider just the first 17 of 87 individually named surgeons. Part of the published analysis comprises a logistic regression on the pooled data without a 'surgeon effect' but including known risk factors for cardiac mortality: the resulting fitted probabilities, when added over a surgeon's cases, give an expected mortality adjusted for the severity of illness of his patients. The ratio of observed to expected mortality can be interpreted as the surgeon's standardized mortality rate, which

TABLE 2

Observed, expected and risk-adjusted mortality for surgeons after CABG surgery, 1990–92†

Surgeon	*Cases*	*Deaths*	*Observed mortality rate OMR*	*Expected mortality rate EMR*	*Risk-adjusted mortality rate RAMR*	*95% confidence interval for RAMR*
Bergsland, J.	613	5	0.82	2.36	1.04	0.33–2.42
Tranbaugh, R.	284	6	2.11	4.11	1.54	0.56–3.34
Britton, L.	447	7	1.57	2.50	1.88	0.75–3.87
Yousuf, M.	433	9	2.08	3.27	1.90	0.87–3.61
Raza, S.	618	12	1.94	2.66	2.19	1.13–3.82
Vaughn, J.	456	9	1.97	2.67	2.21	1.01–4.20
Quintos, E.	259	6	2.32	3.05	2.28	0.83–4.95
Ferraris, V.	276	9	3.26	4.06	2.40	1.10–4.56
Bennett, E.	257	6	2.33	2.50	2.79	1.02–6.07
Foster, E.	266	8	3.01	2.95	3.05	1.31–6.01
Cunningham, J. R.	436	11	2.52	2.47	3.06	1.53–5.48
Bhayana, J.	607	17	2.80	2.61	3.21	1.87–5.13
Lewin, A.	762	19	2.49	2.17	3.43	2.06–5.36
Borja, A.	545	22	4.04	2.69	4.49	2.82–6.81
Canavan, T.	478	19	3.97	2.37	5.02	3.02–7.83
Lajos, T.	636	33	5.19	3.02	5.14	3.54–7.22
Older, T.	222	13	5.86	3.21	5.45	2.90–9.32

†Part of Table 4 of New York State Department of Health (1993), ranked by risk-adjusted mortality rate.

when multiplied by the state average mortality of 2.99% provides a 'risk-adjusted mortality rate' which forms the basis for comparisons between individuals.

Fig. 7(a) shows these ranked risk-adjusted mortality rates with the 95% intervals: with each surgeon's name is shown their risk-adjusted mortality rate expressed as a fraction of their number of cases. Not having access to the patient-specific data, we make the approximation that all a surgeon's patients had the same expected mortality and this leads us to slightly narrower 95% intervals: however, we match the results in New York State Department of Health (1993) by identifying two surgeons as significantly above and one as significantly below the state average mortality. Fig. 7(a) also shows the effect on estimated mortality when assuming that the surgeons are exchangeable with a Gaussian population distribution for logit(RAMR/100), which leads to a more conservative finding that only one surgeon now has an interval that excludes that state average. Estimates are from a simulation with iterations and prior distributions matching those in the previous example.

Fig. 7(b) shows the intervals for the rankings for the independent and multilevel model. It is clear that the intervals are very wide and for the independent estimates we can be confident about whether five surgeons lie in the upper or lower half: Green and Wintfeld (1995) used the fact that 'in one year 46% of the surgeons had moved from one half of the ranked list to the other' to cast doubts on the accuracy of the risk adjustment method, but such variability in rankings appears to be an inevitable consequence of attempting to rank individuals with broadly similar performances. The random effects rank intervals are even more conservative, with only two individuals confidently in the bottom half. In a recent *New York Times* article entitled 'Death-rate rankings shake New York cardiac surgeons', the doctor who was ranked 87th out of 87 in the 1993 tables said 'I want to tell the next poor guy at the bottom of the list not to panic' (Bumiller, 1995).

In parallel with the distinction between adjustment for type A and type B factors in education, a clear difference exists between the role of patient-specific and hospital-specific variables. Further, hospital-specific factors will include both 'structural' variables, such as the number and training of staff, the availability of resources and the throughput of patients, and 'procedural' variables such as the particular operative procedures used. Volume is traditionally associated with improved performance although its association may have been exaggerated (Sowden *et al.*, 1995): Silber *et al.* (1995) illustrated how the variability associated with different groups of factors may be explored and displayed.

5.3. *Interpretation and Impact*

The extent to which even risk-adjusted differences in outcomes can be attributed to the quality of the institution or clinician will always be hotly debated in a context where experimental randomization is not considered to be feasible. McKee and Hunter (1994, 1995) have provided a good discussion from a UK perspective, emphasizing the limitations in availability of data and quality and the difficulty of fully adjusting for context and selective admission policies.

The aim of explicit comparisons is, presumably, to encourage improvements in the quality of care. It is clear, however, that there are several techniques by which the results of such an exercise can be manipulated—this is known as 'gaming' in the USA. An obvious example is provided by Green and Wintfeld (1995) in which the

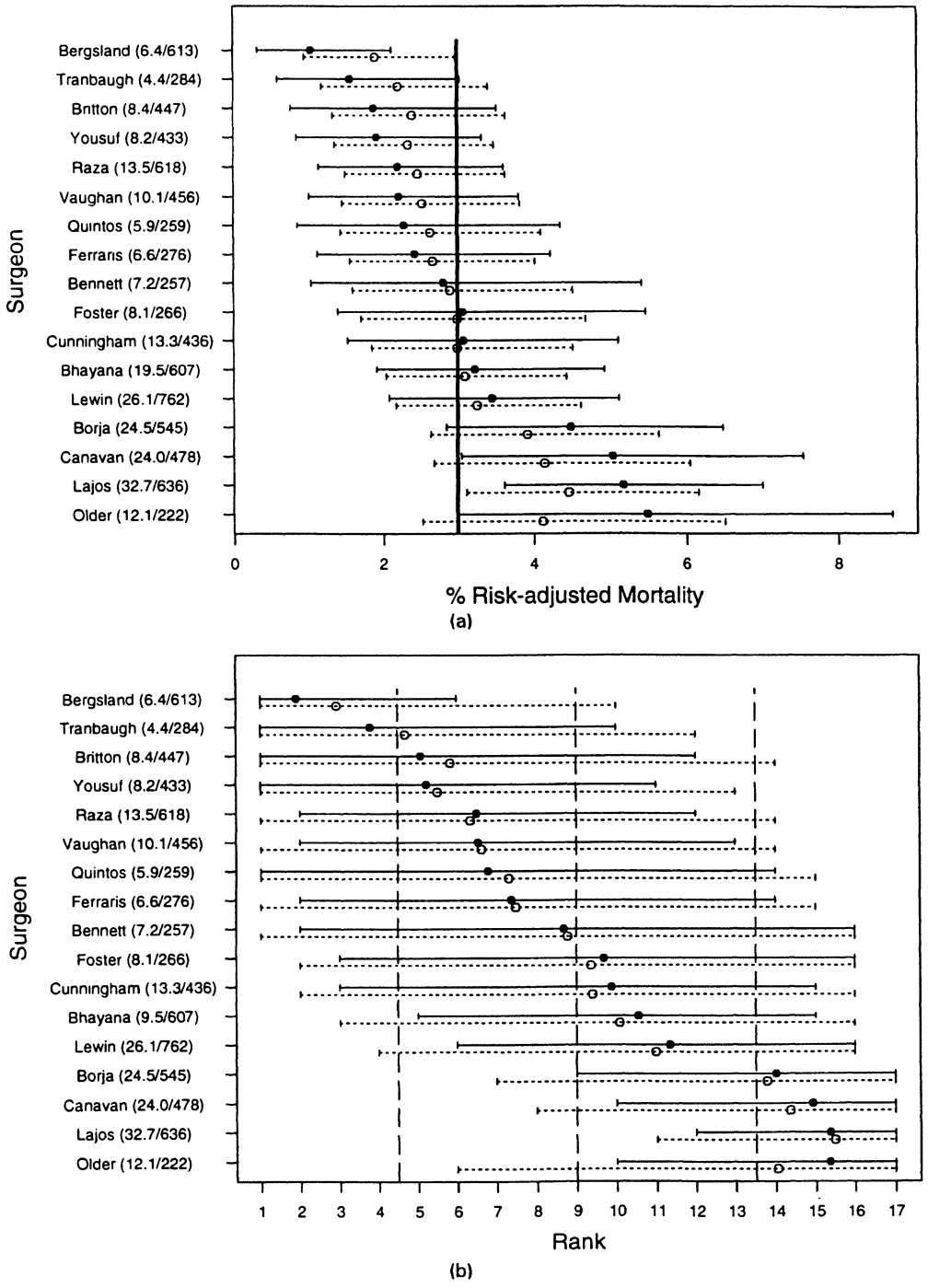


Fig. 7. (a) Estimates and 95% intervals for risk-adjusted mortality rates assuming independent (———) and exchangeable (·····) surgeons (the state average of 2.99% is shown); (b) mean and 95% intervals for rank of surgeon assuming independent (———) and exchangeable (·····) surgeons (fixed effects —●, mean; ⊢——⊣, 95% confidence interval; shrunk estimates —○, mean; ⊢·····⊣, 95% confidence interval)

reported incidence of risk factors that would increase expected mortality rose after the introduction of the programme: for example, reported congestive heart failure rose from 1.7% in 1989 to 7.6% in 1991. McKee and Hunter (1995) pointed out that in the UK the imprecise description of a 'finished consultant episode' allows considerable scope for the inflation of activity, although selective admission of patients and selective reporting of results are other possible strategies for improving apparent performance.

6. CONCLUSIONS AND DISCUSSION

This paper has not treated in detail the issue of the quality and appropriateness of data in both adjustment and outcome measures. Even where available measures are judged to be acceptable, however, there are inevitable limitations in making comparisons between institutions and the paper has concentrated on an exploration of these. Certainly, in our current state of knowledge it seems fairly clear that we should exert caution when applying statistical models to make comparisons between institutions, treating results as suggestive rather than definitive. We have discussed the need for appropriate adjustments and for providing model-based uncertainty estimates. We also need to be aware that for any given set of variables there is often a choice between models, each of which may 'fit' the data equally well, yet give different sets of institutional estimates. This is illustrated in the case of A-level results when total as opposed to average examination scores are used (Department for Education, 1995a).

This implies that current official support for output league tables, even adjusted, is misplaced and governments should be concerned that potential users are properly informed of their shortcomings. If such tables continue to be produced then they need an accompanying warning about their use. Recently, the Department for Education and Employment (Department for Education, 1995a) has published analyses and charts of A-level institutional performance indicators, adjusted for GCSE scores and using efficient multilevel modelling techniques. These analyses recognize differential effectiveness and are based on the results for a complete cohort of students. In our view they constitute an important official move in the right direction. Nevertheless, the continuing official publication and ranking of unadjusted scores lends any comparisons based on them an authority that they do not have.

An overinterpretation of a set of rankings where there are large uncertainty intervals, as in the examples that we have given, can lead both to unfairness and to inefficiency and unwarranted conclusions about changes in ranks. In particular, apparent improvements for low ranking institutions may simply be a reflection of 'regression to the mean'.

A distinguishing feature of many of the outputs discussed in this paper is that they are influenced more by factors that are extrinsic to the institutions than by those for which institutions might be held to be accountable. The identification and measurement of such factors may be very difficult, and it is this feature, predominantly, which makes individual comparisons between institutions difficult to interpret.

Nevertheless, we believe that comparative information about institutions can be useful if handled sensitively with due regard for all their problems, and that this must inform public discussion. There are certain kinds of institutional information which it may be justified to disseminate widely. Information about the physical environ-

ment of a hospital or school or the quality of the organization and management of an institution is relevant to those charged with funding and administering institutions as well as those seeking to use them. Certain kinds of process information may also be useful. For example, the manner in which decisions are reached by the staff of a school or hospital may indicate something important about the quality of life within the institution. There are of course problems with obtaining accurate estimates when samples are small and care will be needed. However, there are some aspects of the process variables which are related to factors over which the institution may have little control. Thus, the exercise of discipline within a school will to some extent depend on the intake social characteristics of the students so that it will be important to make adjustments for this. Likewise, measures of school attendance will often need to be contextualized in terms of social and environmental factors. Furthermore, it is not always clear what aspects of any process variables are 'desirable' — usually we have little detailed information about the relationships between processes and final outcomes. There is also the problem of conveying an appropriate interpretation of uncertainty intervals to the general public and some careful thought needs to be given to this.

In the broad context of resource allocation, information from output indicators, even where valid, may only deserve a small weight. Suppose, for example, that it were possible to identify a school or a hospital which could be held responsible for a relatively poor performance. Suppose further that resources were required to assist such an institution to improve, by way of better management say. Even if this were desirable, there remains the issue about whether any available resources would best be spent in such ways or, for example, on those institutions with relatively poor amounts of input resources, i.e. as part of a policy of 'positive discrimination'. This illustrates the need to consider the use of outcome indicators in a more general context when decisions are taken about matters such as resource allocation.

The examples that we have discussed are concerned with published performance indicators. In some cases, however, systems for the private reporting of indicators have been developed where the results are communicated only to the institutions involved. One such scheme is the A-level information system (Fitz-Gibbon, 1992) which compares A-level results for individual departments within schools and colleges after adjusting for GCSE and other factors. Each institution receives information about its own adjusted residual whereas the remaining institutions are anonymized. Although such systems avoid some of the potential abuse of results which fully public systems can suffer, their inherent secrecy would seem to lend itself to manipulation by institutions, e.g. by ignoring the existence of wide uncertainty intervals or by the selective quotation of results. There would seem to be scope for some important research concerned with the way in which institutions use and respond to performance indicator information, whether public or private.

Finally, although we have been generally critical of many current attempts to provide judgments about institutions, we do not wish to give the impression that we believe that all such comparisons are necessarily flawed. It seems to us that the comparison of institutions and the attempt to understand why institutions differ is an extremely important activity and is best carried out in a spirit of collaboration rather than confrontation (McKee and Hunter, 1995). It is perhaps the only sure method for obtaining objectively based information which can lead to understanding and ultimately result in improvements.

The real problem with the simplistic procedures which we have set out to criticize is that they distract both attention and resources from this worthier aim.

ACKNOWLEDGEMENTS

This work was carried out under grants from the Economic and Social Research Council as part of the Analysis of Large and Complex Datasets Programme. We are most grateful to several referees for their extremely helpful comments.

REFERENCES

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies (with discussion). *J. R. Statist. Soc.* A, **149**, 1–42.

Annals of Thoracic Surgery (1994) Using outcomes data to improve clinical practice: building on models from cardiac surgery. *Ann. Thor. Surg.*, **58**, 1808–1884.

Audit Commission (1995) *Local Authority Performance Indicators*, vol. 1, *Education, Social Services and Total Expenditure*. London: Her Majesty's Stationery Office.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, **10**, 3–66.

British Medical Journal (1995) Hospital league tables derided. *Br. Med. J.*, **311**, 200.

Bumiller, E. (1995) Death-rate rankings shake New York cardiac surgeons. *New York Times*, Sept. 6th.

Charlton, J. R. H., Hartley, R. M., Silver, R. and Holland, W. W. (1983) Geographical variation in mortality from conditions amenable to medical intervention in England and Wales. *Lancet*, i, 691–696.

Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

de Courcy-Wheeler, R. H. B., Wolfe, C. D. A., Fitzgerald, A., Spencer, M., Goodman, J. D. S. and Gamsu, H. R. (1995) Use of the CRIB (clinical risk index for babies) score in prediction of neonatal mortality and morbidity. *Arch. Dis. Chldhd*, **73**, F32–F36.

Department for Education (1994) GCSE and GCS A/AS level performance of candidates attempting two or more GCE A/AS levels in 1992/93. *Statistical Bulletin 9/94*. Department for Education, London.

——(1995a) *GCSE to GCE A/AS Value Added: Briefing for Schools and Colleges*. London: Department for Education.

——(1995b) GCSE and GCE A/AS level performance of candidates attempting two or more GCE A/AS levels in 1993/94. *Statistical Bulletin 4/95*. Department for Education, London.

Department of Health (1994) *Public Health Common Data Set 1993. England*, vol. 1. Guildford: Institute of Public Health.

Fitz-Gibbon, C. (1992) School effects at A level: genesis of an information system. In *School Effectiveness Research Policy and Practice* (eds D. Reynolds and P. Cuttance). London: Cassell.

Gatsonis, C. A., Epstein, A. M., Newhouse, J. P., Normand, S.-L. and McNeil, B. J. (1995) Variations in the utilisation of coronary angiography for elderly patients with an acute myocardial infarction: an analysis using hierarchical logistic regression. *Med. Care*, **33**, 625–642.

Gatsonis, C. A., Normand, S.-L., Hiu, C. and Morris, C. (1993) Geographic variation of procedure utilisation: hierarchical model approach. *Med. Care*, **31**, YS54–YS59.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994) A language and program for complex Bayesian modelling. *Statistician*, **43**, 169–177.

Goldstein, H. (1995) *Multilevel Statistical Models*. London: Arnold.

Goldstein, H. and Healy, M. J. R. (1995) The graphical presentation of a collection of means. *J. R. Statist. Soc.* A, **158**, 175–177.

Goldstein, H. and Sammons, P. (1996) The influence of secondary and junior schools on sixteen year examination performance. In *School Effectiveness and School Improvement*. To be published.

Goldstein, H. and Thomas, S. (1996) Using examination results as indicators of school and college performance. *J. R. Statist. Soc.* A, **159**, 149–163.

Gray, J., Jesson, D. and Goldstein, H. (1995) *Changes in GCSE Examination Performance using Value Added Analysis over a Five Year Period in One Local Education Authority*. Cambridge: Homerton College.

Green, J. and Wintfeld, N. (1995) Report cards on cardiac surgeons: assessing New York State's approach. *New Engl. J. Med.*, **332**, 1229–1232.

Hannan, E. L., Kilburn, H., Racz, M., Shields, E. and Chassin, M. R. (1994) Improving the outcomes of coronary artery bypass surgery in New York State. *J. Am. Med. Ass.*, **271**, 761–776.

Iezzoni, L. I. (1994) (ed.) *Risk Adjustment for Measuring Health Care Outcomes*. Ann Arbor: Health Administration Press.

Jencks, S., Daley, J., Draper, D., Thomas, N., Lenhart, G. and Walker, J. (1988) Interpreting hospital mortality data: the role of clinical risk adjustment. *J. Am. Med. Ass.*, **260**, 3611–3616.

Leyland, A. H. and Boddy, F. A. (1995) Measuring performance in hospital care: the example of length of stay in gynaecology. *Eur. J. Publ. Hlth*, to be published.

Leyland, A. H., Pritchard, C. W., McLoone, P. and Boddy, F. A. (1991) Measures of performance in Scottish maternity hospitals. *Br. Med. J.*, **303**, 389–393.

Lowry, S. (1988) Focus on performance indicators. *Br. Med. J.*, **296**, 992–994.

McArdle, C. S. and Hole, D. (1991) Impact and variability among surgeons on postoperative morbidity and mortality and ultimate survival. *Br. Med. J.*, **302**, 1501–1505.

McColl, A. J. and Gulliford, M. C. (1993) *Population Health Outcome Indicators for the NHS: a Feasibility Study*. London: Royal College of Physicians.

McKee, M. and Hunter, D. (1994) What can comparisons of hospital death rates tell us about the quality of care? In *Outcomes into Clinical Practice* (ed. T. Delamothe), pp. 108–115. London: British Medical Journal Press.

——(1995) Mortality league tables: do they inform or mislead? *Qual. Hlth Care*, **4**, 5–12.

McPherson, K., Wennberg, J. E., Hovind, O. B. and Clifford, P. (1982) Small-area variations in the use of common surgical procedures: an international comparison of New England, England and Norway. *New Engl. J. Med.*, **307**, 1310–1314.

Morris, C. (1983) Parametric empirical Bayes inference: theory and applications (with discussion). *J. Am. Statist. Ass.*, **79**, 47–65.

National Health Service Executive (1995) *The NHS Performance Guide 1994–1995*. Leeds: National Health Service Executive.

National Health Service Management Executive (1992) *The Health of the Nation: a Strategy for Health in England*. London: Her Majesty's Stationery Office.

——(1993) *Local Target Setting—a Discussion Paper*. Leeds: National Health Service Management Executive.

New York State Department of Health (1993) *Coronary Artery Bypass Surgery in New York State, 1990–1992*. Albany: New York State Department of Health.

Normand, S.-L., Glickman, M. E. and Gatsonis, C. A. (1995) Statistical methods for profiling providers of medical care: issues and applications. *Report HCP-1995-1*. Department of Health Care Policy, Harvard Medical School, Boston.

Orchard, C. (1994) Comparing healthcare outcomes. *Br. Med. J.*, **308**, 1493–1496.

Organisation for Economic Co-operation and Development (1992) *Education at a Glance*. Paris: Organisation for Economic Co-operation and Development.

Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*. London: Institute of Education.

Raudenbush, S. W. and Willms, J. D. (1995) The estimation of school effects. *J. Educ. Behav. Statist.*, to be published.

Rogers, W. H., Draper, D., Kahn, K. L., Keeler, E. B., Rubenstein, L. V., Kosecoff, J. and Brook, R. H. (1990) Quality of care before and after implementation of the DRG-based prospective payment system. *J. Am. Med. Ass.*, **264**, 1989–1994.

Rowan, K. M., Kerr, J. H., McPherson, K., Short, A. and Vessey, M. P. (1993) Intensive Care Society's APACHE II study in Britain and Ireland—II: outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *Br. Med. J.*, **307**, 977–981.

Sanders, W. L. and Horn, S. (1994) The Tennessee Value-Added Assessment System (TVAAS): mixed model methodology in educational assessment. *J. Persnnl Eval. Educ.*, **8**, 299–311.

Scottish Office (1994) *Clinical Outcome Indicators—1993*. Edinburgh: Clinical Resource and Audit Group.
Silber, J. H., Rosenbaum, P. R. and Ross, R. N. (1995) Comparing the contributions of groups of predictors: which outcomes vary with hospital rather than patient characteristics? *J. Am. Statist. Ass.*, **90**, 7–18.
Smith, P. (1990) The use of performance indicators in the public sector. *J. R. Statist. Soc.* A, **153**, 53–72.
Sowden, A. J., Deeks, J. J. and Sheldon, T. A. (1995) Volume and outcome in coronary artery bypass graft surgery: true association or artefact? *Br. Med. J.*, **311**, 151–155.
Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1995) *BUGS: Bayesian Inference using Gibbs Sampling, Version 0.30*. Cambridge: Medical Research Council Biostatistics Unit.
Steedman, J. (1980) *Progress in Secondary Schools*. London: National Children's Bureau.
Thomas, N., Longford, N. T. and Rolph, J. E. (1994) Empirical Bayes methods for estimating hospital-specific mortality rates. *Statist. Med.*, **13**, 889–903.
Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996) Adjusting for measurement error in multilevel analysis. *J. R. Statist. Soc.* A, **159**, 201–212.
Yates, J. M. and Davidge, M. G. (1984) Can you measure performance? *Br. Med. J.*, **288**, 1935–1936.

DISCUSSION OF THE PAPER BY GOLDSTEIN AND SPIEGELHALTER

The following contributors were invited to lead the discussion.

Rosemary Butler (Department of Health, London): I am grateful to the Society for the opportunity to discuss this important topic. The Department of Health has always published statistics covering a variety of measures of input, process and outcome. The recent move into publishing performance tables reflects an increased emphasis on providing information for consumption by individual members of the public rather than for public accountability or professional consumption. This move flows from the *Citizen's Charter* initiative; the fundamental principle is that performance in the public sector should be measured, and that the public have the right to know how their services are performing. There is also a business agenda; publication acts as a lever for change and a spur to better performance. Goldstein and Spiegelhalter's paper focuses on technical issues but is light on the context, which is of vital importance, and which explains many of the technical problems to which they refer. Goldstein and Spiegelhalter argue for contextualization of indicators; I argue that the use of the indicators must be seen in context first.

I agree that simplistic interpretations of league tables can be misleading; the question is, can they be avoided given the audience, and does any harm done seriously detract from the benefit of publishing the information? Surely Goldstein and Spiegelhalter would not suggest that we suppress such publications? If the information for the public is to have impact, it must be simple to understand. I agree that information on limitations and uncertainties should be provided, but how can this be done so that the public will understand? Can such measures be sufficiently simple to be used by the public without being derided by professionals for oversimplicity? Subtle statistical techniques will not always help—where tables are produced for the public and interpreted by journalists, any subtle or difficult footnotes or interpretations will tend to be deleted, ignored or not read. Simple messages on uncertainty may be translated by the press as 'these tables are worthless'. Even professionals do not always read everything: in their enthusiasm to criticize the star ratings on our performance tables for a lack of uncertainty measures, Goldstein and Spiegelhalter appear to have overlooked the detailed statistical notes included with the tables, which provide just this information! The introductory text sets out the definitions of significant increases and decreases, i.e. changes of plus or minus 5 percentage points.

Whatever the shortcomings of such performance tables, they have been effective, and performance in the areas covered (which the public regard as important) has improved. They provide indicators of performance which do just that—indicate not define performance; they identify those who have some explaining to do, and those who may have some good practice to spread. Of course the indicators are not perfect, and if most of the apparently poor performers can explain their performance in terms of other important factors of context then the indicator has failed. However, we should not allow the inevitable vociferous criticism by those who can justify alternative explanations of apparently poor performance to outweigh the effect of exposure on those who cannot.

The choice of indicators is often as important as their use. In health, the relationship of health with intervention is often complex with many influences over many years; outcome is often a cumulative end point of interventions by many institutions in the context of wider influences. This was the thinking behind the discussion of population health outcome models and indicators in the 1993 consultation document (Department of Health, 1993). The indicators show what is possible given current data and act as a prompt for the development of relevant indicators in the future.

What should be the Society's role? I suggest that the Society is uniquely placed to draw on the expertise of its members in a wide range of areas, in both the public and the private sectors.

We may wish to consider what makes a good performance indicator and whether there are common messages from different fields. In health we might look at the following.

(a) What aspects of performance are covered by the indicator?
(b) What are the objectives of performance in the context of the indicator?
(c) What standards of service are required to achieve the objectives?
(d) How does the indicator relate to the objectives?
(e) Is the indicator well defined?
(f) Are data of suitable quality available?
(g) How are the data to be presented?
(h) What guidance is needed to help the target audience?
(i) How do we interpret the indicator?
(j) How useful is the indicator likely to be?
(k) What potential is there for follow-up action leading to change?

Are there key pieces of information (e.g. uncertainty measures) which should be made available with performance indicators intended for public consumption and with those intended for professional consumption? Will these pieces of information affect the interpretation by the public or professionals (the public, for example, will always want to know who is top and who is bottom even though the National Health Service performance tables do not themselves give rankings)?

There are some particular areas for investigation in health.

(a) Outcome indicators are most useful if their values can be clearly attributed to the behaviour of the organizations being compared—but how best can we do this in a non-experimental setting?
(b) What are the best methods of presenting change in the context of the base-line, for management and public consumption?
(c) What methods can be used to analyse composite indicators?

Are there messages for us elsewhere?

The purpose of indicators is to indicate performance. The publication of indicators leads to improved performance by stimulating informed questioning on apparently good or bad performance. We need to understand and work within the limitations of the indicators to meet the business as well as the public and the technical agenda. If we wait for the perfect indicator we shall wait for ever. There is a need to balance the good and bad aspects of the use of indicators. The Society has much expertise to offer. But, to have the impact that it deserves, it will need to address the business agenda on choice and use of indicators as well as the technical issues.

Carol Fitz-Gibbon (University of Newcastle upon Tyne): One topic is how we use indicators.

I am grateful to a colleague, Alan Colver, from the medical faculty at Newcastle for giving me these measles data. When he started to give feed-back to health clinics, Northumberland was among the worst, but after a few years of feed-back it was among the best. He did not publicize Northumberland health clinics as failing clinics; nor did he offer performance-related pay or pay per rate of immunization. It was simply feed-back to professionals that others were doing better—and it was effective. Thus we have a hypothesis about how we use indicators: as information fed back to those who are responsible. If that alone works, it is much cheaper than embedding indicators in systems of performance-related pay, threats and so on.

I agree with the authors who have worried about the manipulation of data. If there is fear in the system, there is a constant effort to manipulate the figures.

In the school performance tables that are about to be published, the percentage of A–C-grades has

become a key indicator. All over the country, schools are targeting students who are currently likely to be awarded a D-grade. We have only to look at a very recent publication from the Secondary Heads Association in which Heads are describing how they have improved their schools. Several have even put it in print that they have obtained money from a training and enterprise council to target the D-grade student — this is the new 'special needs' student, for which there has been no legislation. I think that the Audit Commission should 'blow the whistle' here.

The unit of analysis, the unit of aggregation and the unit of reporting are all major issues. I think that the unit of analysis in most of the performance tables and in much research on school effectiveness is wrong, in that account must be taken of major features in the data before deciding to aggregate. Using a procedure called the Kelly–Lawley correction factors as the basis of relative ratings — which is based on an article in 1976 by Alison Kelly and an appendix by Lawley, and which Scotland was using a long time ago — an iterative procedure can be used to balance out who is doing what subject and how difficult each subject is, taking as the indicator of the ability of the students how they do in all the subjects. This is not done correctly unless it is solved taking account of which students are taking which subjects. A rank ordering of subject difficulties can then be obtained that is very similar to the rank ordering obtained by looking at the intercepts from the different regressions (Fig. 8). If each A-level subject is regressed against average General Certificate of Secondary Education (GCSE) score, each regression line will have a different intercept and a different slope.

This feature of the data must be modelled. To aggregate to the school level is to lose all the useful information within the school. It is not even much use to parents. They do not want to know how good the language department is, if their child will be doing mathematics and science after 16 years of age; they want to choose on the basis of a good mathematics and science department.

What matter are the variables that are measured. In our work on value added for the School Curriculum and Assessment Authority we have looked at many multilevel models, and also at simple ordinary least squares — what might be called 'residual gain analysis'. One regression is put through the entire set of pupil level data — everyone in the school can understand that more or less — and then the residuals for each school are gathered together. This is accessible, and well understood by the users. The residuals obtained in this way have not yet correlated less than 0.94 with the residuals from multilevel modelling — but try to explain multilevel modelling to an entire staff!

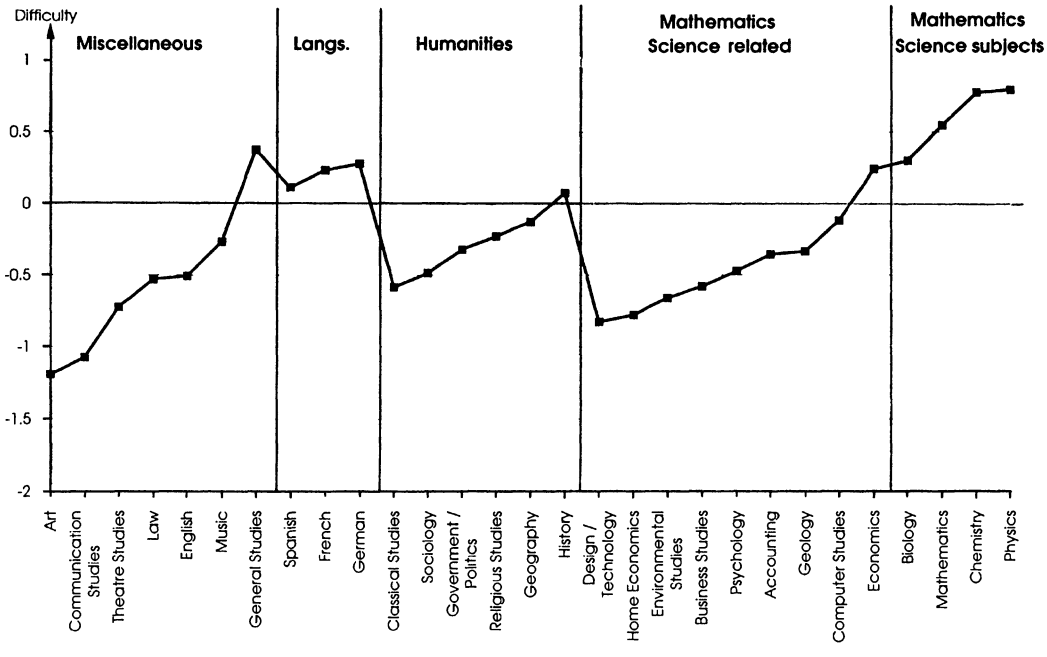


Fig. 8. Correction factors for 1993 (source: ALIS, candidates who took at least two A-levels)

If we look at the correlations with the *average* GCSE as opposed to the *total* GCSE score, early on (in 1983) it was clear that the average GCSE score was the correct input measure for looking at the A-level results subject by subject—and that analyses had to be subject by subject both to be of any use to schools and to understand what was going on.

The reason why the total GCSE score is not such a good predictor is that, confounded in the total, are school policies. In some schools, students can take only eight GCSEs; in others 12. It is the level of performance that predicts A-level grades not the school policy about the entries. I am happy that after 12 years the Department for Education and Employment has now shifted from the total to using the average.

I would like now to turn to a more fundamental problem with school effectiveness research and with these indicators, which is that all the data are passive observational. They are *not* causal data and cannot be directly interpreted as causal data. I am delighted that the Centre for Evidence-based Medicine has opened at Oxford. I feel that we should be hearing this message more strongly in educational research. To quote David Sackett

> 'If you are scanning an article about therapy and it is not a randomized trial, why are you wasting your time?'.

The more we work with totally confounded passive observational data, the more frustrated we become; it serves only to confuse, it does not tell us what works, what we can do or what is the way forward.

It is not helpful, however, when prestigious people make statements implying that there is no room for experimentation in education. Education is *the* compulsory treatment: thousands of people are being treated every day with very few treatment withdrawals. It is the most exciting possible social science. Is it moral to keep dealing with schools in certain ways instead of systematically trying to find out better ways? There is no morality in *not* using experimentation in social science; it just leaves us ignorant, and there is no morality in ignorance. There are many totally ethical ways to run experiments, such as equal funding, but different ways forward. I find schools very interested in randomized controlled trials at multiple sites. I do not hear this message coming from the statisticians, and I think that it needs to come from them—and try to impress it on the politicians.

We might wonder whether very complicated models have to be used. Are many variables needed in the models? If adjustment is made for reliability, there is a reduction in the various residuals. If adjustment is made for the intake, that is when there are the changes in the rank ordering. To add socioeconomic status variables then hardly changes things.

This is at the A-level stage; it will be slightly different before then. The correlation with home background at A-level is about 0.1, and before A-level about 0.3—so it accounts for about 9% of the variance. It is nothing compared with the 50% of the variance accounted for by prior achievement.

I am saying that extremely simple models are possible, simple ordinary least squares or, as far as schools are concerned, drawing the line of best fit and seeing where the pupils lie around it. We also encourage them to use statistical process control charts. In these Shewhart charts it is entirely arbitrary where the confidence limit lines are drawn. The traditional 5% level has been raised out of all sense of proportion. Where does it come from? Apparently Fisher could not obtain permission for the other levels, so he published at the 5% level—and now there is an entire generation testing everything at the 5% level!

Finally, if we do not have proper statistics and do not involve the users in their use, we are leaving schools with statements like

> 'The school promotes satisfactorily the social and moral development of the pupils, but not their spiritual and cultural development'

(from a report by the Office for Standards in Education) being used to judge them. I would like to know the error terms on the judgment of 'social' and 'moral', and how can they be right on those and wrong on the 'spiritual' and 'cultural'? That failed a school. The head teacher had been there for 19 years and was retiring. His retirement was shattered; he had no come-back; he could not challenge this in court. How was that judgment made?

John Gardner (Department for Education and Employment, London): Though I admit to being one of the slow learners from the Department for Education and Employment (DFEE) I am trying to improve, and I can take some credit for the move to average A-level scores from total A-level scores in the modelling research that the Department published in the summer of 1995. I would like to say how

much we welcome the paper by Goldstein and Spiegelhalter. It is a landmark—I am not aware of any earlier work where statisticians in health and education have collaborated so fruitfully with each other.

I share Carol Fitz-Gibbon's wish for some experimental research in education. It is extremely rare and very difficult to bring about, because there are always accusations of social engineering if any attempt at experimental design is made, for example if pupils are randomly allocated to schools, even for research. Without such experimentation, however, it is very difficult to interpret data from observational studies, and of course this problem is partly what the paper is aiming to address.

To some extent the two groups of statisticians contributing to the paper have conducted an experiment, because health and education are very similar in many ways: they are both public services, national services locally administered, where responsibilities are devolved to local units. In both cases it is difficult to decide exactly what is a good outcome from the treatment either in hospital or school, and it is therefore inevitable that there will be a range of uncertainty in the measurements made and problems with both their reliability and their validity.

It is also true—and I think that this is a particularly intractable problem—that the outcome for any patient or student is strongly influenced by the input into the process brought by the patient or student themselves. This effect is very important, and difficult to allow for—as we saw in the 'gaming' example that was mentioned in the US study.

The paper in fact reports on some kind of experiment. Two sets of statisticians, in health and in education, have been put to work in these different areas, without initially conferring, and have come up with their respective statistical methods. The paper starts to explore the differences between them. I would welcome more work in this direction because I am still not certain, perhaps because I do not fully understand the methods, why multilevel modelling is appropriate for education while risk adjustment is suitable for health. They are obviously similar approaches, but there are differences; it would be helpful to tease them out and to show how they might affect the kinds of conclusion reached.

An important point raised by the paper, but which has not been discussed so far, is the difference between type A and type B adjustments for context. The distinction between them can help to explain the apparent differences of view between the researchers and the position taken by the government. In my view, the consumers of the performance tables (not of course 'league' tables) are primarily parents and students aged 16 years who, I think it is generally acknowledged, would be interested solely in type A adjustments: they want to know how the particular institution will help them to reach their desired level of achievement, given their starting point. They are not bothered by the managerial efficiency or internal organization of the school, so that many of the adjustments made at the type B stage are of limited interest to them.

I certainly think that users of the tables are also interested in the raw figures. Like Rosemary Butler, I would defend them. Parents need to know what the average pupil in a given school or college has achieved at the end of the course, in the hope that their own child will achieve a similar level.

The need for accuracy in the published figures has not been mentioned, though it is absolutely paramount. Much effort is expended by the DFEE in checking the examinations and other data to ensure accuracy. Even if the figures may not be relevant to user needs, accuracy is important in its own right because it is essential that schools can recognize and take ownership of the published figures. They must be able to follow and agree their derivation. The Department devotes much effort and expense on the collection process, using two very able contractors in FORVUS and Bath University to help to produce and check the figures.

The education tables have certainly proved useful to parents. We have conducted some qualitative research, particularly of parents of pupils around 10 years of age—the age when pupils generally move to their secondary school. Wide awareness and acceptance of the tables was found, with about three-quarters of all parents aware—perhaps hardly surprising given the huge volume of information given in all the national and local newspapers when the tables are published. About 50% of parents claimed that the tables were useful to their choice of secondary school and about 10% said that the tables had had a major effect on this choice.

This strong influence of the tables might be rather worrying, because we all acknowledge that there are many factors other than examination results to take into account. Indeed, the research shows that other factors are valued by parents, among them the proximity of the school to their home and the general reputation and ethos of the school—will the parents, and the children of course, feel comfortable in that particular school? The research shows that these factors have an important weight.

The paper is of course mostly concerned with how to estimate uncertainty. Section 4.2.1 shows how uncertainty intervals might be derived from the raw figures, giving education as an example. Evidently

the researchers did not have full access to pupil level data, although they do exist and it might be interesting to explore their use. Whether uncertainty measures should be shown in the tables or not raises several issues, one of which is that when parents use the figures to make their choice they have to take into account various factors other than the examination results. It is arguable that the uncertainty intervals applying to one particular statistic or factor may be of limited interest. Parents have to use some kind of implicit Bayesian model, partialling in all the various factors that have to be taken into account, of which examinations are only one. Perhaps if the researchers can tell us how to do that, we might think harder about how to present uncertainty in the tables.

We all agree that value-added measures would probably be more useful to parents than the raw figures but, as we all know, value-added estimates are not available yet.

The paper illustrates an important point about the construction of a school improvement index. It shows that changes over time between schools are even more difficult to estimate than the comparison of a point in time. This is one reason why, I think with considerable regret, the Department decided not to publish an improvement index in this year's tables—not because we thought that the index was inherently flawed as a concept, but because it seems to be unmeasurable, certainly with the available data.

Finally, I hope that we could think more about how uncertainty could be presented in the tables for parents without confusing them too much, and given that many people say that the tables are already overloaded.

Alison Macfarlane (National Perinatal Epidemiology Unit, Oxford): The use of statistics to assess institutional performance dates back to well before 1840 when the Statistical Society of London set up a Committee on Hospital Statistics (Statistical Society of London, 1842, 1844). Papers about the interpretation of hospital statistics appeared in the early journals of the Statistical Society (Guy, 1856; Steele, 1861, 1877; Burdett, 1882), raising questions which are still important and unresolved.

In the third edition of her *Notes on Hospitals* (Nightingale, 1863), Florence Nightingale noted that

> '. . . accurate hospital statistics are much more rare than is generally imagined and at the best they only give the mortality which has taken place in the hospitals and take no cognizance of those cases which are discharged in a hopeless condition, to die immediately afterwards, a practice which is followed to a much greater extent by some hospitals than others'.

She went on to comment about the need to adjust for the type of patients admitted to a hospital. She also suggested that measures of hospital morbidity were needed:

> 'Careful observers are now generally convinced that the origin and spread of fever in a hospital or the appearance and spread of hospital gangrene, erysipelas and pyaemia generally are much better tests of the sanitary state of a hospital than its mortality returns'.

Goldstein and Spiegelhalter have focused largely on methods in their discussion of 'statistical issues'. I would suggest that, as in the past, the Society's concern covers the whole process of collection, analysis and publication of relevant data.

Collection is a key issue. Most of the data cited by the authors have been chosen either because they are collected already or because they can readily be collected. Thus the data used in the *NHS Performance Guide* (National Health Service, 1995), waiting times for out-patient appointments and in-patient admissions, operations cancelled, the percentages of operations done as day cases and operations cancelled, are measures of process rather than outcome. They are free-standing indicators without links to information about the characteristics of the patients, the severity of their illness or the outcome of the care that they received. It is not known whether the pressure to reduce reported waiting times affects the way that data are collected or whether the publicity given to what is measured can have detrimental effects on unmeasured aspects of health care.

Many of the so-called 'population health outcome indicators' (McColl and Gulliford, 1993) are also measures of activity rather than outcome. They were chosen from data assembled in the public health common data set, instead of collecting new data to answer questions about the health of the population. They are a fragmented set of counts and rates rather than a complex data set with record linkage to relate outcome to treatment and the people treated.

Questions arise about whose performance is being measured. The authors assumed that teenage conception rates reflected the performance of health boards. They did not appear to consider whether they also reflect the characteristics of the population, the activities of the clergy or the quality of personal and social education given in school.

Classification and analysis of data pose further questions. The authors mention briefly, but probably underestimate, the problems involved.

While supporting the authors' use of interval estimates I question their assumption that ranking data is the most important aim of analysis. In the National Health Service 'performance tables' hospitals and trusts are grouped with varying numbers of stars. Although the raw data provided allow an explicit ranking, there are many ties. In any case, it may be more useful for service providers and users to know how their own figures compare with those for similar areas and the national average or median (Audit Commission, 1995).

For analytical techniques to be useful in the public arena, it is necessary for them to be understood. If I, as a statistician, find some of them difficult to follow, I wonder how community health council members would fare, let alone the tabloid press.

The correct interpretation of data can often be problematical, particularly if politicians gain the upper hand. For example, government statisticians make a clear distinction between 'in-patient episodes' counted and patients treated, but Ministers have an interest in blurring this distinction. Recent moves to giving government statisticians a greater role in presenting their data to the media and others are therefore welcome. It is also important to ensure that there is independent and well-informed comment and criticism and the Royal Statistical Society should have a role here.

More generally, what should we do? The authors' paper is a reflection of the divide between theory and practice in statistics. Theoretical statisticians have a worrying tendency to develop methods and then to look for data, whereas applied statisticians need appropriate techniques to answer practical questions. The Society should do all that it can to bridge this divide. This meeting and Harvey Goldstein's considerable involvement in public debate on school league tables are positive signs.

Mary Smalls (National Health Service in Scotland, Edinburgh): My first reaction on reading the earlier version of this paper was indignation at finding the Clinical Resource and Audit Group outcome indicators described as league tables. Great care was taken within the National Health Service (NHS) in Scotland to discuss the indicators within the service before they were made publicly available, and to load the publication with *caveats* and assumptions: indeed most of what is identified in Section 5.1 of the paper. Every internal publication of the indicators has a government health warning—'no inferences should be made on quality of clinical care from this data'—and we refused to publish the data in any sorted order. So why have the authors included them in a discussion of league tables?

Informing the debate

An example of the positive aspects of our publication would be that, when a medical director was surprised at his position on one of the indicators compared with that of his peers, he investigated further, and found a consistent error in clinical coding. This is now being corrected and the quality of the data improved.

The authors are correct to highlight the complexities of measuring institutional effectiveness, but it is even more complex to define an outcome in health terms—especially when there are no readily available measures of inputs to the process, such as severity of disease on admission. However, it should not be forgotten that we are dealing with observational, rather than experimental, data. Defining an appropriate specification for the underlying statistical model when there may be 20 or more covariates, some of which will not even have been identified, requires some deep data dredging. In the absence of relevant, reliable, measurable indicators health service managers adopt proxies, which they think that they can understand, usually in only one or two dimensions.

This point is important because I do not believe that it is possible to discuss the statistical issues in isolation from the business issues. Chatfield (1995) highlighted the importance of focusing on the problem to be solved, and the interpretation and communication of results, rather than dredging the data to find the best fitting model.

How can the Society contribute to the debate?

Management scientists and accountants, including the National Audit Office, are well ahead of the game and probably see little need for a contribution from statisticians, especially when they want to add to the complexity of the situation rather than to reduce it. For example, I find Fig. 6 on ranked residuals poorly explained in the paper. Public health experts would look at the data presented and say fine, there seems to be no problem here since the islands have a small population; we cannot be confident on the ascertainment of the data because we know that teenage pregnancies can be hidden by staying with relatives during the confinement, when religion plays an important part both within the family and in the attitude of doctors; and what is a better measure of the effectiveness of health promotion?

This suggests that we are not interested in statistically significant differences but in addressing the questions: is there cause for concern?; how can we provide feed-back to staff to encourage them to improve their practice?; can we trigger further investigation and action to improve the quality of inputs, process and outputs? A considerable amount of energy is currently being spent on good practice benchmarking. Here the aim is not regression to the mean, but to increase the quality of processes by adopting the best of practice followed by similar groups. I think that if the Society wishes to influence the direction of institutional performance indicators then it cannot do so from the academic high ground, but it should follow an approach closer to that of Smith (1990).

Although I have been critical of the academic treatise of the subject, I do welcome new methods of estimating confidence intervals for population data. It has taken us at least 15 years to get customers to understand interval estimates rather than point estimates. Extracts of the Scottish linked database have been made available to colleagues in Glasgow University, who are applying multilevel models. We await applications which are robust, can be understood by professionals, academics, the press and other stakeholders, and provide added value to the tax-payer in understanding the performance issues in the health service, if the extra processing costs can be justified.

David Draper (University of Bath): I have several comments on this interesting paper.

Quality assessment: input–output versus process

The quality of hospital care (e.g. Kahn *et al.* (1988)) is usually thought to involve three things: process (what providers of care do on behalf of patients), measured implicitly (with the physician's judgment) or explicitly (with objective, if-then-else criteria); outcomes, such as mortality (what happens to patients as a result of process); sickness at admission, since any judgment of appropriateness of hospital outputs must take account of inputs. League tables work with the last two of these ingredients but ignore the first — this is a kind of *input–output* (I–O) analysis, in which what goes on inside the black box (process) is inferred (at best) indirectly.

To learn about the validity of quality assessment information provided by I–O analyses, it is instructive to take process P as a tentative standard and to see how well I–O replicates P. I have analysed data from a major quality-of-care study in the USA (Kahn *et al.*, 1990) in which process, mortality and sickness at admission were measured on an average of 35 patients sampled randomly from each of 297 representative hospitals. To serve as 'truth' for this analysis, I identified hospitals as 'actually bad' or 'actually good' if they were respectively below or above the pth percentile of the distribution of average process scores across the 297 hospitals; to simulate I–O screening results I labelled hospitals as 'apparently bad' or 'apparently good' if they were respectively below or above the pth percentile of the distribution of z-scores for excess mortality after adjustment for admission sickness. I cross-tabulated these two dichotomies to estimate the positive and negative predictive values (PPV and NPV) of I–O screening as a function of p, with results as in Table 3. To assess the effect of sample size within institutions, I sorted the 297 hospitals on average process and merged adjacent hospitals on this list, to yield 148 first-stage 'hospitals' each with a median of 69 patients, and then merged adjacent hospitals again to yield 74 second-stage 'hospitals' each with a median of 138 patients.

To interpret a typical value in Table 3, e.g. with $n = 69$, if you rank ordered hospitals on excess mortality and then did a detailed process audit on the apparent 'worst' facilities, of the 20% sample

TABLE 3

Estimated positive and negative predictive values (PPV, NPV) of I–O hospital quality screening†

% of hospitals audited	*297 hospitals (median n = 35)*		*148 stage 1 'hospitals' (median n = 69)*		*74 stage 2 'hospitals' (median n = 138)*	
	PPV	*NPV*	*PPV*	*NPV*	*PPV*	*NPV*
10	0.20 (0.07)	0.91 (0.02)	0.20 (0.10)	0.91 (0.02)	0.25 (0.15)	0.91 (0.04)
20	0.31 (0.06)	0.83 (0.02)	0.40 (0.09)	0.85 (0.03)	0.47 (0.13)	0.86 (0.04)
30	0.39 (0.05)	0.74 (0.03)	0.45 (0.08)	0.77 (0.04)	0.48 (0.10)	0.76 (0.06)
40	0.47 (0.05)	0.65 (0.04)	0.51 (0.07)	0.67 (0.05)	0.50 (0.09)	0.66 (0.07)
50	0.58 (0.04)	0.57 (0.04)	0.59 (0.06)	0.59 (0.06)	0.62 (0.08)	0.62 (0.08)

†Values in parentheses are binomial standard errors.

tentatively identified by the I–O screen as among the 20% worst, 40% would be among the 20% worst as measured by process. Given that a purely random screen would have achieved a 20% PPV under these conditions, the estimates in Table 3 are disappointing (even accounting for substantial sampling variability) and reinforce the authors' results on the substantial width of confidence intervals for ranking institutions.

Optimal measurement strategy?

The I–O approach to hospital quality screening is at the (low cost, low accuracy) end of a continuum whose (high cost, high accuracy) end point is explicit process; for example, abstraction times per record based on the Kahn *et al.* (1990) study ranged from 20–30 min for mortality adjusted for admission sickness up to 100–120 min for explicit process. If you had, say, £100 million to spend on measuring quality of hospital care in the entire UK in a given year, how much should you spend on process and how much on adjusted outcomes? This is the key measurement question, and almost nothing is known about its optimal answer. It would seem from the low PPVs in Table 3 that if you want to know what is going on inside the black box you have to open it—after all, even if an I–O analysis has tentatively identified a substandard institution, you will still need to measure process to figure out what is wrong—but there may yet be some cost-effective role for I–O screening if properly applied.

'Random' effects with non-random data

All the models in the paper treat the institutional effect as 'random', even though there is nothing random about the institutions involved: we typically have data on *all* the units (students or patients) at *all* the institutions (schools or hospitals) in, for example, some particular geographical area, *at one moment in time*, and our interest is in *these* institutions, not in any hypothetical population from which we might pretend these institutions were 'randomly' drawn. What therefore justifies the use of 'random effects' models in this case?

One answer is that our interest is typically in saying something about *future units at these institutions*—we are hoping that the device of random effects, which yields shrinkage estimates of underlying mean outcomes, will produce better predictions next year than treating the institutional effects as 'fixed'. There has been little predictive validation to support this hope; more would be welcome. In this context it is worth noting that predictive improvement from shinkage depends on time homogeneity of the process under study. For example, Bill Browne and I have recently been looking at admission rates for children to a Bath hospital from each of the 124 wards near Bath in 1991–94. In this data set both fixed and random effects predictions for 1992 based on 1991 are poor: across the 124 wards only 76% (fixed) and 75% (random) of the nominal 95% intervals cover, because there was a big (about 20%) overall rise in admission rates from 1991 to 1992. Shrinkage is of little value when the process under study is changing in ways that are not captured by your random effects model.

Alternative models and priors

In Section 5.2.3 the authors assume that 'the surgeons are exchangeable with a Gaussian population distribution for logit(RAMR/100)', leading to the shrinkage results in their Fig. 7(a). Other models for the data of Table 2 are perhaps more natural; for example, using the logit of the expected mortality rate as a severity-of-illness measure s_i for surgeon i, who has r_i observed deaths in n_i patients, one might plausibly take $(r_i|\pi_i) \sim \text{bin}(n_i, \pi_i)$, $\text{logit}(\pi_i) = \alpha + \beta(s_i - \bar{s}) + \theta_i$, $(\theta_i|\sigma_\theta^2) \sim N(0, \sigma_\theta^2)$, $\alpha \sim N(0, \text{huge})$, $\beta \sim N(0, \text{huge})$ and $\sigma_\theta \sim U(0, c)$. This model produces results that differ by 10–20% from those reported in Fig. 7(a); for example, for surgeon Bergsland the authors' model produces the 95% shrinkage interval (1.1, 3.0) but the model above gives (0.7, 2.3). Discussion with Dr Spiegelhalter indicates that the only difference between the paper's model and the model above is that the authors take β to be 1.0, rather than using the data to assess plausible values of β. In fact 1.0 is not particularly strongly supported by the data—the posterior mean and standard deviation of β are 0.64 and 0.68 respectively (these figures incidentally call into question the quality of the severity-of-illness measures used by New York State to produce the expected mortality rates). As for priors, with a small number of level 2 units (e.g. surgeons), various ways of trying to specify a 'diffuse' prior for the level 2 heterogeneity parameter σ_θ can lead to somewhat different answers; for example, the BUGS manual and examples for version 0.30 favour a gamma(ϵ, ϵ) prior for the precision $\tau_\theta = 1/\sigma_\theta^2$. This back-transforms to a prior on σ_θ with a large amount of mass near 0, which requires justification case by case. With little prior information about heterogeneity I prefer a $U(0, c)$ prior on σ_θ, with c chosen to span the range over which the marginal likelihood for σ_θ is appreciable. Results using this prior with the surgeon data again differ from those in the paper by about 10%.

Causal inference

Statistical adjustment is causal inference in disguise; recasting the problem in causal terms can be clarifying. For example, in Section 3.1 the authors say that

> 'If [aspects of institutions and general practitioners] are influential then they should be incorporated into the statistical model, if trying to estimate an effect associated with institutional performance'.

This is not necessarily true: just because a confounder Z is strongly associated with an outcome Y it does not automatically mean that Z should be an adjustor in a model trying to estimate the causal effect of a third factor X on Y—it depends on the appropriate counterfactual (Draper, 1995). For instance, the presence or absence of a computer tomography (CT) scanner at a hospital is strongly associated with good outcomes for stroke patients, but it is irrelevant to estimate what the outcomes would have been with a CT scanner at rural hospitals for which there will never be enough money to buy such a machine.

Rex Galbraith (University College London): Harvey and David have provided a very good basis for discussion of statistical issues concerning institutional 'league tables'. I particularly liked the emphasis on uncertainty and the interesting consequence of using the more precise 'shrunken' estimates of performance, namely that the corresponding ranking is *less* precise. This might be regarded as an argument against shrinking, but to my mind it is an argument against ranking. I shall comment briefly on ranking *versus* performance measure, fixed *versus* random effects and graphical presentation.

Harvey and David imply that institutions will inevitably be ranked. But we should emphasize the performance level rather than the rank. A high rank does not guarantee a high performance level. Furthermore there will typically be a few institutions that stand out at each end of the scale and relatively large numbers in between with similar performances that cannot reasonably be ranked.

Of course the performance measure should be meaningful. Comparing risk-adjusted mortality rates after cardiac artery bypass graft surgery seems a more convincing exercise than comparing average A-level scores for various schools. Knowing a school's average A-level score tells us very little of use, particularly when the within-school standard deviation is about 9. Potential users should want to know something much more specific. Members of the institution being assessed need a measure that more directly reflects their performance. I would also like to see the relevant 'error' components of variance established from background research rather than estimated from each new year's data.

It is difficult to display uncertain estimates graphically while preserving both simplicity and integrity. Plotting intervals is an effective method of comparing estimates with a reference value but is less effective for comparing estimates with each other. Fig. 7, although good of its kind, probably devotes too much space to displaying lines. Ironically Fig. 7(b), where the intervals are so wide that one's eye is drawn to the sequence of ordered point estimates, gives the opposite message to that intended. Table 4 shows an alternative version. This is not intended to be a definitive league table, but rather to raise some points of methodology.

(a) The graph area is reduced and the numerical information is emphasized.
(b) Surgeons are ordered by number of cases, not by mortality rate. This is a departure from convention, but it can be more informative to structure the order by a different criterion than the performance measure. Consequences in Table 4 are that the column of numbers of deaths is more interesting, and the intervals are ordered approximately by width.
(c) Interval estimates (not point estimates) are presented for both mortality rate and for rank, forcing the reader to accept the uncertainty.
(d) The graph shows both 50% and 95% intervals; variants are possible.
(e) The graph is plotted with respect to a non-linear scale, so that the precision (or shape of the distribution) does not depend on the mortality rate. Presenting non-linear scales should often be considered, particularly for rates and proportions, and can result in a simpler interpretation.

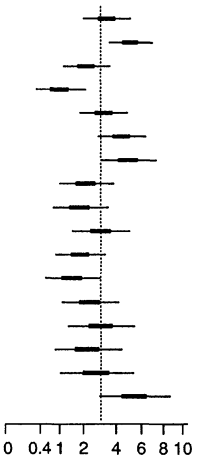
Not all these choices are appropriate for all occasions; there is scope for studying the effects of various forms of presentation.

Another graphical method is to plot the precision explicitly on one axis. Fig. 9 shows a 'radial plot' (Galbraith, 1988, 1994) of the teenage conception data given in Fig. 6. This type of plot was designed for use in a different context, but it is perhaps useful here

TABLE 4

Numbers of cases and deaths and risk-adjusted mortality rates (95% interval estimates) for 17 surgeons, and 95% intervals for rank of surgeon†

Surgeon	*Cases*	*Deaths*	*Risk-adjusted mortality per 100 cases*	*Rank*
Lewin	762	19	2.04–5.13	4–16
Lajos	636	33	3.56–6.99	11–17
Raza	618	12	1.15–3.56	1–14
Bergsland	613	5	0.34–2.13	1–10
Bhayana	607	17	1.87–4.89	3–16
Borja	545	22	2.85–6.38	7–17
Canavan	478	19	3.01–7.36	8–17
Vaughan	456	9	1.01–3.85	1–14
Britton	447	7	0.78–3.48	1–14
Cunningham	436	11	1.51–5.04	2–16
Yousuf	433	9	0.86–3.26	1–13
Tranbaugh	284	6	0.55–2.93	1–12
Ferraris	276	9	1.09–4.18	1–15
Foster	266	8	1.34–5.41	2–16
Quintos	259	6	0.84–4.41	1–15
Bennett	257	6	1.02–5.34	1–16
Older	222	13	2.92–8.68	6–17



†The graph shows 95% and 50% intervals compared with the state average of 2.99 deaths per 100 cases.

(a) for looking at several estimates together and
(b) for seeing the shrinkage effect.

The independent estimates (full circles) are only slightly overdispersed with respect to Poisson variation about a common rate. The open circles are empirical Bayes estimates using a gamma mixing distribution with mean 8.37 and index 55.4, and hence a coefficient of variation of only 13%. They are similar to the exchangeable estimates in Fig. 6 and the distribution of ranks simulated from the joint posterior distribution is practically identical. Fig. 9 shows both the shrinkage and the increase in precision for each health board, and shows how two estimates may change order (e.g. Orkney and Highland). The order of empirical Bayes estimates can also depend on which other points are included; for example omitting Glasgow changes the ordering of Orkney, Borders and Arg&Cly, though their actual rates are similar.

Although there is a strong statistical case for publishing shrunken rather than raw estimates, there may be some opposition from administrators or users. The Bayesian argument is familiar. From a non-Bayesian viewpoint it may be of interest that 'best linear unbiased predictors' were developed in the context of animal breeding as a technique for estimating genetic merits, particularly for selection and ranking (e.g. Robinson (1991)).

A. Eastwood, T. A. Sheldon and P. Smith (University of York): The authors are to be congratulated on the paper which discusses the technical aspects of performance analysis with thoroughness and elegance. We wish to discuss more contextual issues and to emphasize both the behavioural and the economic aspects of this area. Smith (1996) suggests three stages in the assessment of performance: measurement, analysis and action. The authors have provided an excellent insight into analysis; we wish to address the equally important stages of measurement and action.

A performance indicator is an attempt at providing an unbiased measure $\hat{\theta}$ of some area of measurable activity M which is thought to be a general reflection of overall performance P:

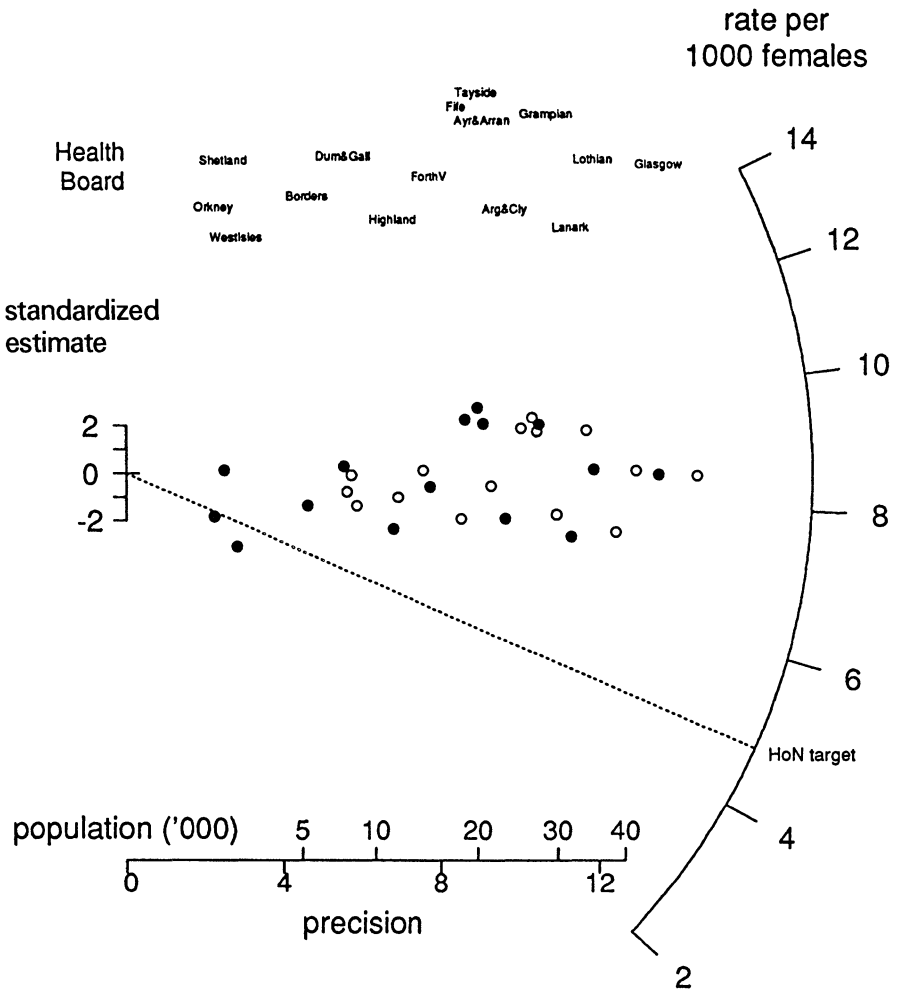


Fig. 9. Radial plot of the teenage conception data: ●, rates based on an independent Poisson model (fixed effects); ○, empirical Bayes estimates using a gamma mixing distribution (random effects)

$$\hat{\theta} = f(M),$$
$$P = g(M, O)$$

where O is all other measurable and unmeasurable aspects of performance. Implicit in this model is the assumption that M and O are positively correlated. In contrast, if we assume that an organization is operating efficiently, economic theory would suggest that M and O are likely to be traded off against one another and so an improvement in the measured activity M can only be gained at the expense of a deterioration of the unmeasured activities O. Organizations rewarded for performing well on M will devote resources to improving activity M at the expense of O (Kerr, 1975; Smith, 1995). This may result in unintended effects of performance measurement such as the reduction in the level of performance of activities in O which may reduce the overall level of P. Thus, although $\hat{\theta}$ is still an unbiased estimate of M, it ceases to be a good measure of overall performance P.

Another effect of measuring M is that the organizations will start to manipulate or game the system to give a false impression of high levels of performance. Thus, for example, in coding hospital episodes diagnostic groups may be shifted to appear more severe, in the research assessment exercise the status of

researchers becomes redefined and denominators become changed. In this case, although the relationship between M and P might still hold, $\hat{\theta}$ will now be a biased estimate of M and so give a false impression of performance.

Thus once a measure has been chosen as an indicator of performance it may cease to be an accurate measure of that which is regarded as performance! The distortion of behaviour and the way in which performance ceases to be reflected in the indicators used is illustrated by how the Soviet Union fared when they tried to run the whole economy by using performance measures (Kornai, 1992).

A further consideration is that the information systems required for performance indicators can be quite expensive in terms of collection, analysis and interpretation of data, and this expense needs to be justified in terms of the expected gain. The economic benefits of performance indicators will depend on the relative importance of identifying differences or not mistakenly reporting statistically significant differences between organizations. The paper uses the 5% significance level throughout which implies a certain view about the relative costs of type I and II errors. There is, however, no evidence that 5% is the correct level when seeking to justify good and bad organizational performance.

Usually measures are chosen for administrative convenience or ideological appeal rather than for their ability to capture truly the essential dimensions of performance as defined by workers and consumers. Thus, measurable activity M is unlikely to be highly correlated with P, even if it were to be estimated accurately.

Even if the unintended behavioural effects do not occur, the next issue is how the results are interpreted and affect subsequent action (Smith, 1996). If performance measurement is not clearly linked to the process of work and management then it is not clear how a particular result will trigger an appropriate response. If, for example, coronary heart disease rates in one area are higher than those in another or increase over time, it will be difficult to disentangle the influences of the health services from those of non-health factors (e.g. socioeconomic profile and base-line differences in the stock of health). Even if this is possible, it is not easy to attribute a poor result to a specific health factor (e.g. lack of preventive action in primary care, poor hospital treatment or food policy). If performance cannot be linked to process, then consequent action will not necessarily be appropriate, possibly resulting in unnecessary investments in irrelevant areas of activity. The multilevel methods reported here therefore need to be clearly linked to policy responses.

Thus, although the statistical issues of handling uncertainty and making legitimate comparisons between units are important, of equal significance are the behavioural and economic context and effects of performance management. These are likely to affect both the ultimate validity of the measurement and also the costs and consequences of performance measurement.

The following contributions continued the discussion.

Jane I. Galbraith (London School of Economics and Political Science): The authors and earlier discussants have argued cogently that ranks are unreliable and unwanted (except for publicity), that producing a single performance indicator is inadequate and possibly counter-productive and that appropriate adjustments should be made for confounding factors. I shall therefore confine my comments to shrinking!

The fixed effect estimates (unshrunk) are appropriate for type B analyses, e.g. for examining or explaining variation between institutions. For example a radial plot of school scores standardized by the within-school standard error might show excess variation between schools or one or two outliers. But, for a parent choosing between two or three local schools, a type A analysis, such as interval estimates of the random effects (shrunk), should provide a better guide.

However, the amount of shrinkage may be subject to model bias or estimation error. If the within-institution variance is overestimated then the between-institution variance will be underestimated and the indicators will be excessively shrunk, and vice versa. For example, if clustering had caused extra-Poisson variation in teenage conception rates within health boards, then the Poisson model would not have shrunk the estimates sufficiently. (In fact the Poisson model appears to fit nicely to the data for each health board for the three separate years.) In the example of mean A-level scores for 10 schools, the likelihood is rather flat for a wide range of within and between components of variance. So here there is little to choose between no shrinking of the fixed effects and total shrinking to a common value.

Perhaps the authors will comment on how the interval estimates take into account the estimation error in the within and between components of variance, and what allowance should be made for uncertainty in the model specification.

J. B. Copas (University of Warwick, Coventry): A new and highly contentious area for performance indicators is the probation service. At the top of the list of proposed performance indicators is the comparison of actual against predicted reconviction rates. The predictions are to be made using a prediction score which I developed for quite a different purpose, now to be taken as the base-line for a value-added calculation. Here the base-line is up front, not a backroom adjustment to be added later.

Those who wish to discredit the whole exercise have an easy target: just rubbish the prediction score. This was done with vehemence in the press in the summer of 1995. The arguments advanced showed an all-pervading lack of numeracy and a hostility to the very idea that statistical calculations have anything useful to say about matters involving professional judgment.

There are several major issues for statisticians:

(a) taking statistics out of context—our *caveats* are often ignored;
(b) the population is always changing;
(c) global goodness of fit is not enough—we need to pay much more attention to local variations, perhaps involving random effects models as we have seen here;
(d) the tacit assumption is made that cases are sampled from a population—nearly always there will be non-ignorable screening;
(e) we need a measure of *total* uncertainty, covering not only sampling errors within our models but also errors of misspecification and omission, which are probably much more important.

Ramesh Kapadia (Office for Standards in Education, London): This timely paper raises several important issues on the use of league tables. These comments relate to the use of league tables in judging school performance, a central aspect of the work of the Office for Standards in Education.

The term value added, which relates to a monetary measure in economics, is certainly difficult to justify in education: inputs and outputs cannot be measured with the same units and there is an implicit assumption that only what is measurable is important. Nevertheless, some indication of success needs to be developed.

It is crucial that there are no basic inadequacies in the appropriateness and integrity of data: thus, although 'raw' league tables mask important contextualizing factors, the data are usually reasonably reliable. Since there is undoubted media interest in these tables, ensuring that the data are reliable becomes paramount: this is generally accepted with regard to data on examination performance, but problems with data on attendance remain.

Data in adjusted league tables are much more open to question, particularly if the data are confounded in relating to both individuals as well as institutions. The extent to which reliance can be placed on rankings in adjusted league tables requires further discussion, particularly where only contextual rather than prior attainment data are used.

Changing the variables measured can also make substantial differences; though the convention in research is to use point scores with equally spaced weightings for grades A–G in the General Certificate of Secondary Education, it is more common in schools and the media to use thresholds, and to give proportionally greater weighting to the higher grades.

An interesting distinction is made between 'type A' comparisons (in choosing between institutions with regard to performance) and 'type B' comparisons (measuring characteristics which lead to the differences). The major focus is on type A comparisons; it would be valuable to learn more of research on type B comparisons, which includes such issues as the vexed question of class size.

The paper's view that current official support for league tables is misplaced requires further analysis, particularly the examples which are thought to lead to unfairness or inefficiency, when the tables are given an undue degree of validity. Accountability will remain an issue of concern; it is likely that output indicators will be used in determining resource allocation.

There is scope, as suggested, for systematic and careful research on the way that institutions use and respond to performance indicator information. In education, the publication of league tables has had some effect in schools, particularly in urban areas. The aim must be to obtain objective information which can lead to better understanding and improvement.

Ian Schagen (National Foundation for Educational Research, Slough): First let me say how much I welcome the authors' paper and their commitment to the proper use of performance indicator data in both education and the health service. I shall confine my comments to education. Although the authors rightly reject many of the current abuses of league tables, they seem to retain an interest in comparing

schools with each other and in producing rankings, albeit with qualifications. However, I do not believe that this is the most useful function of school level performance indicators.

A sophisticated analysis of a school's results is most valuable when fed back directly and confidentially to that school's management, to be used in combination with other information in a process of school improvement. For this, schools need comparisons with the mean, with what might be expected of a school of their type with their sort of pupils. Of more importance, in many ways, than overall performance indicators are subject level analyses which can help them to pinpoint departments or curriculum areas which are performing above or below expectations.

One of the outcomes of the National Foundation for Educational Research's 'Quantitative analysis for self-evaluation' initiative has been the identification of departments producing results significantly above those expected in certain schools. Qualitative research on the features of such departments was able to identify consistent common factors, which could then be fed back to other schools and departments to help with their own self-improvement process. Details of this work were reported at the 1995 European Conference on Educational Research at Bath (see also Harris *et al.* (1995)). This kind of analysis is ultimately more useful to educational improvement than the comparison of school with school in some kind of 'league table', however sophisticated.

R. Brand (Leiden University): I shall concentrate on two problems which the authors have noted and show that possible solutions exist, by using the annual quality comparison of all Dutch obstetrics clinics as an example.

This on-going project uses routinely collected data; each centre annually receives its own observed and expected mortality rates over the past 5 years, the latter based on the results of all centres and adjusted for risk factors that hospital policy should not be held accountable for. The analyses are based on about 400 000 live-births and rank all clinics on the difference between observed and expected mortality. Reports are distributed anonymously through a notary's office.

The authors quote

> '. . . the fact that "in one year 46% of the surgeons had moved from one half of the ranked list to the other" casts doubts on the accuracy of the risk adjustment method'.

Indeed, no matter how judiciously the models are fitted, yearly mortality rankings are highly variable. Therefore we never provide yearly significance tests but use five-year cumulative ranks instead. No ranks should be issued when the between-years correlation is extremely small. We follow the *trend* in ranking and emphasize the *level* and *direction* in which they develop.

In 1995 we started to provide obstetricians with adjusted rankings on the incidence of major interventions: Caesarean sections, induction of labour and assisted deliveries; the main purpose was to provide them with adjusted data, primarily focused on 'introspection' rather than on 'confrontation'.

The authors warn us that

> '. . . institutions and individuals will seek to improve their subsequent ranking'

and

> 'An overinterpretation . . . can lead both to . . . unwarranted conclusions about changes in ranks'.

To avoid this we created a unique situation: nearly all Dutch obstetrics clinics are randomized and only 50% will be informed about their ranking on 'tendency to interventions'. In 1998 we can analyse the effect of the reporting systems. By simultaneously monitoring the neurological status of the newborns, we guard against a possible overreaction of clinicians, which could result in an overall adverse effect on health outcome.

In conclusion we argue that

(a) major ranking projects should not be undertaken or financed for a period less than 5 years,
(b) conclusions on rankings should always take into account correlations over time and
(c) analogously to the severe conditions rightfully imposed on the introduction of any drug or treatment in society, no ranking system should ever be introduced and used by health authorities nor by the profession itself without a true randomized trial to prove the efficacy of the entire procedure and the absence of any major adverse effects.

Thomas A. Louis (University of Minnesota, Minneapolis): The beauty of the Bayesian formalism is its ability to structure complicated models, inferential goals and analyses. The prior and likelihood produce

the full joint posterior distribution and it generates all inferences. Applications have burgeoned now that Markov chain Monte Carlo methods enable relevant models.

Unfortunately, the authors fail to take full advantage of this Bayesian formalism. When the goal is ranking (schools, physicians or geographic regions), the analysis must directly infer ranks. An indirect approach can have a very poor performance. For example, unless the posterior distributions are all from the same symmetric location–scale family with identical scale parameters, ranking posterior means produces inappropriate ranks. Furthermore, even when the posterior variances are all equal, the optimal estimate of the empirical distribution function (EDF) of the co-ordinate-specific parameters (the ensemble) is not the EDF of the posterior means.

As Ghosh (1992), Laird and Louis (1989), Louis (1984) and Tukey (1974) showed, posterior means shrink too far and so are too closely clustered around the prior mean. Laird and Louis (1989) proposed ranking methods based on posterior expected ranks and, importantly, associated posterior (confidence) intervals and applied their approach to the Aitkin and Longford (1986) school effectiveness data. Goldstein and Spiegelhalter's analysis of educational performance indicators was protected by the near constancy of the posterior variances, which allowed ranks based on posterior means to produce good ranks. Nevertheless, to give an accurate portrayal of spread and spacing, in displays posterior means should be replaced by the ensemble (constrained) estimates of Louis (1984) and Ghosh (1992).

In the health example, posterior variances differ substantially and it is important to display ensemble estimates and to compute expected ranks. The use of BUGS saved the day, because the authors produced ranks from parameters simulated from the posterior distribution thereby avoiding the mistake of ranking posterior means. They note with some surprise that the ranked posterior means do not mimic the mean ranks, but this discrepancy is unsurprising with unequal posterior variances.

Good unit-specific estimates need not produce good estimates of the parameter ensemble or produce good ranks. However, communication will be aided considerably by a 'triple-threat' set of estimates with an EDF that is a good estimate of the parameter EDF, with induced ranks that are good estimates of the parameter ranks and with good performance in estimating unit-specific parameters. For exponential families with conjugate priors, ensemble estimates perform well (Devine *et al.*, 1994), but additional research is needed.

Alastair H. Leyland (University of Glasgow): This paper suggests that cross-classifications of institution by area of residence may be of importance when constructing indicators of performance or outcome, and I would like to emphasize this point. The data that I have used are related to the recent publication of a table of deaths within 30 days of admission to hospitals in Scotland with primary diagnosis of acute myocardial infarction between 1990 and 1993 (Scottish Office, 1994) and individual patient records are taken from the linked computerized system of Scottish morbidity records (Kendrick and Clarke, 1993). The basic hierarchical model for the response (1 ≡ death) of individual i from area j and hospital k, y_{ijk}, is assumed to be

$$y_{ijk} \sim B(1, \pi_{ijk})$$

with two alternatives for the modelling of the probability of death, π_{ijk}, given the age and sex of each individual (x_{pijk}):

$$\text{logit}(\pi_{ijk}) = \sum_{p=1}^{P} \alpha_p x_{pijk} + u_k \tag{4}$$

and

$$\text{logit}(\pi_{ijk}) = \sum_{p=1}^{P} \alpha_p x_{pijk} + v_j + u_k \tag{5}$$

where $v_j \sim N(0, \sigma_v^2)$ and $u_k \sim N(0, \sigma_u^2)$.

The estimate of σ_u^2 from model (4) is 0.040, but from model (5) $\hat{\sigma}_u^2 = 0.035$ and $\hat{\sigma}_v^2 = 0.181$; the implications of this are that not only are small area effects of more consequence than hospital effects but also that the omission of the area level from the hierarchy will lead to an inflation of the estimated between-hospital variance.

The second point that I would like to make concerns the choice of variables for which an outcome is adjusted. It will seldom be the case that age and sex alone provide a satisfactory explanation of individual or institutional differences. To illustrate this point consider the posterior residual estimates shown in Fig. 10. Two estimates (and confidence intervals) are shown for each of the 15 largest hospitals; the first estimate refers to a model which adjusts for age and sex alone whereas the second considers the additional effect of the secondary diagnosis. The rankings produced by the two sets of estimates would not be the same, and differences between the two become more pronounced when smaller hospitals are considered. There is less precision around the estimates for the model including secondary diagnosis, and the estimate of σ_u^2 increases to 0.071. Clearly the inclusion of more explanatory variables is likely to affect the hospital residuals further, and it is important that influences beyond the control of a hospital are removed from a measure of performance. Neglecting the influence of geographical differences or failing to adjust for confounding variables are just two ways in which the publication of league tables can amount to misinformation.

Sheila Gore (Medical Research Council Biostatistics Unit, Cambridge): I wish to discuss performance indicators in relation to prisons, and suggest that before performance indicators are visited on institutions they should be first used to evaluate policy.

The particular prisons policy I want to discuss is random mandatory drugs testing (MDT), introduced into seven pilot prisons in England and Wales from February 1995. Prisoners who test positively or refuse to provide a urine sample are punished—loss of up to 28 days' remission. The MDT policy has been dubbed 'war on drugs' and promoted as a means of gathering information. As a means of gathering information it is unethical, because it is coercive, the supposed objective would be better served by unattributable urine samples and urine samples cannot distinguish men who have taken drugs by injection—a dire health hazard—rather than orally.

The prison service proposes to base performance indicators for drugs misuse on its MDT programme, in particular, to monitor the proportion testing positively for cannabis and for class A drugs. It ought also to monitor assaults on staff and other prisoners—because violence in prison is associated with drug misuse.

Is the pilot study adequate to evaluate policy? Seven male establishments yield 300 tested samples monthly, about 30% being positive for cannabis and 4% for class A drugs. If the war on drugs is very successful, the proportion of prisoners taking cannabis might be reduced by a third. If this were the case, it would have been known within 2 months. If the policy has a modest effect on the numbers taking cannabis, a reduction by a sixth would be identifiable within 8 months, i.e. *already*.

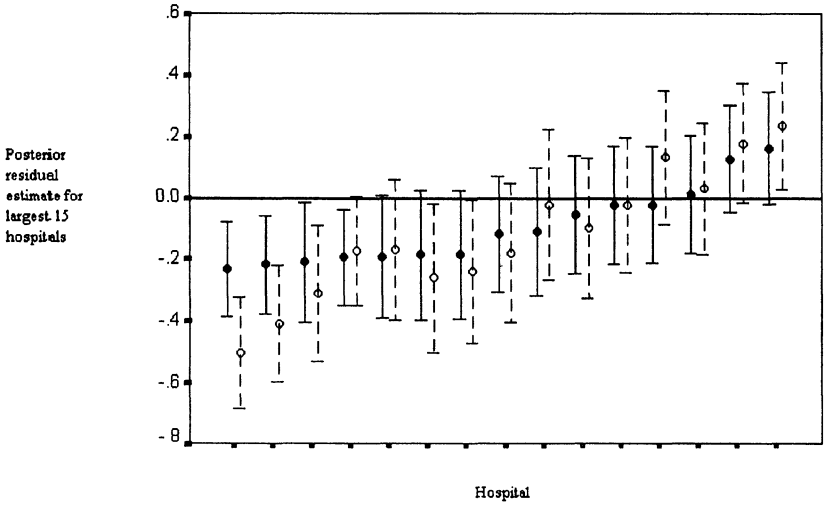


Fig. 10. Paired estimates of posterior residuals (together with 95% confidence intervals) for the 15 largest hospitals under models which include (○) and exclude (●) the effect of secondary diagnosis

However, we also want to make sure that the proportion testing positively for class A drugs is reduced, not increased (14 days half-life for cannabis, 3 days for heroin). Does the war on drugs convert prisoners from oral cannabis to heroin by injection to avoid detection? The pilot study is indeed sufficiently large to detect over a year whether there has been a 50% increase in the proportion testing positively for class A drugs—up from 4% to 6%.

In the seven pilot establishments 400 assaults are expected in a year. There has been no change in the number of assaults over the last 2 years nationally and so, if the war on drugs is successful, a historical comparison would give about 60% power to detect a 15% reduction in assaults.

Prison governors and the public should know the effect of the random MDT policy before it is visited on all institutions.

The performance of performance indicators can be evaluated by simple design criteria—and should be more often.

Gerald Goodhardt (City University Business School, London): I would like to refer to two simple ways in which league tables can be misleading for what might be called 'statistical' reasons.

The first concerns the way that the size of the institution affects the measurement of the criterion variable. In the lists of the best performing unit trusts which appear from time to time in the financial pages of the newspapers, a disproportionate number of smaller trusts is often found. From this it was inferred that small trusts do better than larger ones. All sorts of reasons for this were put forward, such as managers of small trusts are more actively involved and can respond more quickly and flexibly to market changes. The late Sidney Benjamin pointed out that the worst performing trusts also contained a disproportionate number of smaller trusts, and that, on average, small and large trusts performed as well as each other, largely in line with the market. The variation between smaller trusts was greater than between larger trusts, probably for statistical reasons. Another example of smaller institutions dominating both ends of a league table is to be found in Colombo and Morrison's (1988) entertaining paper on submission rates of doctoral theses.

The second problem arises from 'regression to the mean' which is briefly mentioned in the paper. In my field of marketing, it is quite common for a manufacturer to rank the regions of the country by the *per capita* sales of his product in a particular year. He may then decide to do something about what he sees as weakness by increasing advertising in the regions at the bottom of the list, funding this by reducing advertising in the successful regions. When he measures the sales in the following year, lo and behold, the weaker regions where advertising was increased have improved, whereas those in which advertising was reduced have fallen back, thus 'proving' the power of advertising once again!

Turning briefly to the paper, there seems to be an indication of a regression to the mean effect in Figs 2 and 3, making the difference between schools in Fig. 3 even less significant!

Stephen Senn (University College London): This interesting paper is a fine example of the way in which the Society, through learned contributions of Fellows, can play a part in raising the level of public debate on matters of national importance. Shrinkage estimators, however, may not be greeted with the same enthusiasm outside this Society as within and I think that it is important to appreciate that choosing to adopt them might have certain curious, and ultimately political, implications.

For example, the authors' model implicitly assumes that the size of a school has no effect on its performance. (This is not to be confused with whether size affects shrinkage: it does.) This is slightly ironic when we consider the heated debate about whether the size of *classes* (admittedly for rather younger pupils) is an important determinant of academic performance. Consider also two schools with the same mean total A-level scores: one is small and the other is large and both have excellent raw results. The small school will have its results shrunk closer to the overall mean. This is only right and proper from one point of view but from another it can easily be presented as quite unfair. I presume also that the authors would recommend to the government that if they use league tables for schools they should shrink results. What advice, however, would they give those in the universities admitting students on the basis of A-levels? Given two students with identical scores from two different schools, should one, other things being equal (and ignoring a value-added red herring) prefer the student who comes from the school with the better average score on the grounds that her shrunk results will be better? (Of course, we do not measure A-levels twice so the model might be a little difficult to apply but given a little help from Bayes and General Certificate of Secondary Education scores something could be attempted.) I suspect that the authors and other Fellows of this Society would show less enthusiasm for shrinkage estimators in such a case and it is not difficult to think of other circumstances in which

they might be very unpalatable indeed: there are cases, after all, where a prior is scarcely distinguishable from prejudice. Therefore, admirable though this work is, there is a further political aspect which requires very careful consideration by the authors and by other Fellows of this Society before the results of this paper can be turned into a recommendation.

Ian H. Langford (University of East Anglia, Norwich) **and Alastair H. Leyland** (University of Glasgow): Goldstein and Spiegelhalter produce a table of mean rankings, with 95% confidence intervals, for 17 doctors performing coronary artery bypass surgery, using a fixed effects and a multilevel model with random effects between doctors. We also analysed the same data using a random effects model with the MLn software (Rasbash and Woodhouse, 1995), with observed cases O_i and 'risk-adjusted' expected values E_i so that

$$O_i \sim \text{Poisson}(\mu_i)$$
$$\ln \mu_i = \ln E_i + a + b_i$$

where $b_i \sim N(0, \sigma^2)$, a is a fixed constant representing the mean relative risk of mortality and the natural logarithm of the expected values are fitted as an offset, centred around their mean to aid convergence. We then estimated the residuals b_i from this model, using penalized quasi-likelihood with a second-order Taylor series approximation for the residuals (Goldstein, 1995a). We applied two approaches to generate confidence intervals for the rank order of the doctors, namely

(a) a Monte Carlo simulation of 999 values of the 17 residuals for each doctor based on the estimated mean and standard deviation of each residual, assuming normality, i.e. using the 'plug-in' estimates—a delta method approximation is used which adjusts for the sampling variation of the parameter estimates (Goldstein, 1995b)—and
(b) a parametric bootstrap technique (Efron and Tibshirani, 1993), where we simulated residuals (b_i^*) for each doctor from the estimated random effects between doctors. We then added these residuals back onto the fixed part of the model, including the offset, and generated a new set of O_i^*. The model was then refitted, and new residuals b_i' calculated, and the bootstrap distribution of $b_i^* - b_i'$ obtained by repeating the process 999 times. The values of $b_i + b_i^* - b_i'$ were then ranked for the 999 replications to produce a distribution of ranks for each doctor.

Table 5 shows the results from these replications, giving the mean rank and confidence intervals for each technique. Obviously, the Monte Carlo technique is much quicker to implement, as the model does

TABLE 5
Mean rankings with 95% confidence intervals for the 17 surgeons

Doctor	*Monte Carlo ranking*			*Bootstrap ranking*		
	Mean	*−95%*	*+95%*	*Mean*	*−95%*	*+95%*
Bergsland	3.2	1	13	2.8	1	9
Tranbaugh	4.7	1	13	4.8	1	12
Yousuf	5.6	1	14	5.3	1	13
Britton	5.9	1	14	5.6	1	14
Raza	6.4	1	14	6.2	1	13
Vaughan	6.8	1	14	7.1	1	15
Ferraris	7.2	1	15	7.4	1	15
Quintos	7.5	1	16	7.3	1	15
Bennett	8.6	2	16	9.0	1	17
Foster	9.2	2	16	9.4	2	16
Cunningham	9.5	2	16	9.6	2	16
Bhayana	10.3	3	16	10.3	2	16
Lewin	10.8	4	17	11.0	3	16
Borja	13.8	6	17	13.7	6	17
Older	13.8	7	17	14.1	6	17
Canavan	14.1	6	17	14.3	7	17
Lajos	15.4	10	17	15.4	9	17

not have to be refitted for each replication. The two methods produce approximately the same results, which are quite similar to those reported by Goldstein and Spiegelhalter, who used a Gibbs sampler to obtain their estimates. This is encouraging and suggests that using plug-in estimates may be adequate in situations where ranks are required. However, confidence intervals for ranks tend to be resistant to changes in variance, and this is reflected in Table 5. In fact, the bootstrap estimated the variance for doctors to be 3.5% greater than did the Monte Carlo method.

David Bartholomew (London School of Economics and Political Science): I would like to make an observation on the question of communicating uncertainty to the public at large, which I think is a major challenge to statisticians.

We think that we know about uncertainty, and that when we have added a standard error or a confidence interval to a point estimate we have increased knowledge in some way or other. To many people, it does not look like that; they think that we are taking away their certainties—we are actually taking away information, and, if that is all that we can do, we are of no use to them.

This was brought home to me forcibly when Peter Moore and I appeared before the Employment Select Committee of the House of Commons—which is not a random sample of the population at large. Our insistence that we could not deliver certainties was regarded as a sign of weakness, if not downright incompetence. One may laugh at that, but that is the way it was—and that is what we are up against.

We must persist, and I would like to suggest one way in which it might be done in the league table context. Instead of presenting a ranking, we present a set of ordered categories—as we do for classifying degrees but in a rather more sophisticated way. The number of categories and their coverage would depend on the uncertainties, so the uncertainties are implicit. What we would present in fact are some tied rankings, but how we have tied them and in what numbers would reflect how uncertain we feel about them.

Spiegelhalter made some remarks very much on these lines, on his diagrams. He drew attention to groups where members could not be separated. Often one could only be confident about differences between those at opposite extremes. If we did this kind of thing we might convey something. It is statistically inadequate, but much better than nothing.

Nicholas T. Longford (De Montfort University, Leicester): I agree with the view of the authors that in institutional comparisons ranking of units is inevitable. It is not clear whether ranking is always appropriate. Under uncertainty about the estimated ranks, shrinkage may reduce the risk of large errors in assigning the ranks. This implies a reduction in the differentials of rewards or resources allocated in future. Laird and Louis (1989) proposed an elementary method of estimating the ranks by evaluating $\hat{r}_i = \text{rank}(u_i)$ as $1 + \Sigma_{j \neq i} P(X_i > X_j)$ and the sampling variance of $\hat{r}_i$ using the joint probabilities $P(X_i > X_{j_1}, X_i > X_{j_2})$.

Even if the ranks were established with perfect precision, the incentive-based scheme for allocating resources may make little sense. 'Rewarding' higher ranking units by enhanced resources may be wasteful, when they are already well endowed. The objective should be maximizing improvement of the quality of the system as a whole, under the constraints of limited resources. For that, we need to know *why* institutions differ, but also how their performances are affected by incentives, differential rewards and other outcomes of assessment exercises.

When an assessment exercise is carried out regularly (e.g. annually), it is important to distinguish two sources of instability of the ranks: one due to estimation errors and the other due to a genuine variation in performance. Since the assessment lags behind the period assessed, variation in performance may obliterate the purpose of assessment. Imagine that an institution is being rewarded while it is doing a poor job. . . .

My reading of Table 1 is that the hypothesis of no differences among the schools and of constant between-year differences would not be rejected. Yet the discussion of Figs 2 and 3 contradicts this hypothesis. Somewhere along the lines, the uncertainty about the between-school variation has been forgotten. The 10 schools are not sufficient for a good estimation of between-school variation. One resolution of this problem is to incorporate prior information about variation—either in a Bayesian or a frequentist framework, or informally.

Most of the variables used for institutional comparisons are proxies for 'ideal' variables which cannot be observed. In a typical analysis, we rely on an unchanging association of the proxy to the ideal or underlying variables. 'Gaming' can be interpreted as attempts at undermining this association. In

pursuing this, the validity of the assessment is compromised and irreparable harm may be done to the system as a unit. For an example we may not have to go any further than the universities research assessment exercise.

One undesirable result is the preference for the short-term improvement to the consistent and more profound improvement over a long period of time, with large investment; see Gray and Wilcox (1995) for a detailed discussion of institutional assessment in education.

Ian Diamond and Fiona Steele (University of Southampton): We very much support the use of the techniques to produce league tables described in this excellent paper. Our comment concerns a particular application. In the social sciences we are often criticized by the ethnographers and the anthropologists who say that we do not link in with them sufficiently and that we simply produce a set of statistics which do not represent reality.

We have recently found some results which we report in Steele *et al.* (1996) (on immunization in some rural villages in Bangladesh) very useful for linking in with anthropologists. We had great difficulty explaining multilevel modelling to them, but when we start to work with 'league tables' of this sort we can explain how different villages are working in different ways.

David Spiegelhalter said that we are often interested only in the outliers. In fact, by using league tables, we can find examples of places which are perhaps not outliers but where we want to look for the pathways of influence on why they are not outliers. For example, one particular Bangladeshi village would have been expected to have high levels of immunization, whereas it was down in the middle of the table with quite a large confidence interval. This seemed rather strange, but our colleagues were able to attribute this to a fundamentalist imam. It is possible that in this village mobility of women was extremely limited and so babies were not being immunized. Another example is a village at the top of the league table, which our colleagues could attribute to a very enthusiastic school-teacher.

All these things that we did not know about were not included in the model at the beginning but, by connecting with the qualitative workers, by encouraging the fieldworkers to look further at particular villages and by saying to them that we were surprised that this place was good and that one was bad, we could get people to understand the potential for linking the sophisticated statistical methods with qualitative research.

Sharon-Lise Normand (Harvard Medical School, Boston): The authors have presented an important exposition of the statistical issues involved in comparing institutional performance. I shall comment on two themes: the type and role of adjustors employed in multilevel modelling and the choice of performance measure on which to compare institutions.

The authors assume that, having made suitable adjustments, institutions are exchangeable. How do the authors define *suitable* and do they believe that there is room for unadjusted outcomes? There are two levels of adjustors: patient specific and institution specific. Several methodological difficulties exist with incorporating either type of covariate into the multilevel model. For example, physicians may comment more frequently in the medical charts at tertiary care institutions than at small community hospitals. The endogeneity of the recorded information will make it very difficult to separate patient case mix from the institution effect and this will affect the performance measure. Instrumental variables could play some role but it will be difficult to instrument severity and hospitals.

The inclusion of patient-specific covariates to adjust outcomes for risk is an accepted practice; the inclusion of institutional-specific covariates deserves thought. The effect of patient level covariates on the outcome are often not exchangeable across institutions. To maximize the precision of the estimates it seems reasonable to include some structural characteristics of the institution, such as hospital size. However, inclusion of last year's mortality rate when comparing current mortality rates will lead to another endogeneity problem. This argument extends to the patient level model. Patient adjustors are often restricted to those covariates that are unrelated to treatments received when comparing operative mortality even though their inclusion would improve mortality prediction. To infer institutional performance based on non-treatment parameters, treatment-related covariates could be included in estimation and then averaged out.

The choice of a performance measure is central to the issue of comparing and improving performance across institutions. This would argue for performance measures that are directly linked with subject-matter considerations (Normand *et al.*, 1996) and *not* ranks. The length of time to administer a drug on patient entry is a meaningful quantity that clinicians can strive to improve. Another important set of performance measures can be derived from medical guidelines (American College of Cardiology and American Hospital Association, 1990).

In conclusion, policy recommendations based on the results of institutional comparisons will strongly affect the delivery of education and health. I look forward to the development of innovative methodological approaches for this form of analysis.

Martin McKee (London School of Hygiene and Tropical Medicine): Goldstein and Spiegelhalter have, reasonably, focused on the statistical issues related to league tables. I would like to expand on two important issues that they mentioned only briefly.

The first is that of risk adjustment. It may never be possible to eliminate all confounders. In a recent study of outcome following gastrointestinal haemorrhage, adjustment for routinely collected variables changed hospital rankings but further adjustment for variables identified after endoscopy changed the rankings again (Rockall *et al.*, 1995). In a comprehensive review, Iezzoni (1994) has identified 11 dimensions that have been shown to be associated with risk, including cultural factors and patients' preferences. If information is available on all these we could generate almost as many league tables as we want, each with subtly different rankings.

The second related issue is scope for manipulation. Enthusiasm for league tables is generating an impressive anthology of imaginative responses that subvert underlying policy objectives, from recorded crime statistics to train timetables. The intrinsic uncertainty surrounding diagnosis and treatment (Sackett *et al.*, 1991) offers endless possibilities for opportunistic behaviour. In health this has been compounded by using administrative terms with unreal properties of meaning anything one chooses them to (Clarke and McKee, 1992). It only requires that appropriate incentives for manipulation are developed, such as league tables. As noted above, it may be impossible either to detect or adjust for this behaviour.

But what can be done about these concerns? Addresses to learned societies are insufficient because of the political context within which publication is taking place. Methods of presenting information are not value free. In an examination of the role of government in Britain since 1979, Jenkins (1995) has shown how league tables have been used as a weapon in the struggle to achieve an unparalleled concentration of power in the hands of Ministers. The simplification of complex issues is seen as an important method of controlling those whose power is based on discretion in the face of uncertainty (Lancet, 1995), such as clinicians and teachers. Researchers seeking to challenge the uncritical application of league tables must understand the strength of those who see technical arguments as a distraction from their main aims. Unfortunately, these arguments are often complicated and public understanding of statistics is poor, as indicated by the success of the national lottery (Persaud, 1995). The scientific community has a responsibility to educate and communicate. The government and popular press are unlikely to do it for us.

The following contributions were received in writing after the meeting.

Keith R. Abrams and Paul C. Lambert (University of Leicester): We would like to congratulate the authors on a clear and concise presentation of some of the important statistical issues involved in the comparison of institutional performances.

We would also like to point out the close links with other areas in which comparisons or estimations of individual unit effects are required, as well as aggregating in some way their effects to obtain a population or pooled estimate. Two such areas in particular are multicentre clinical trials and meta-analysis. In both settings, though we may ultimately be interested in estimating an overall effect, comparisons of individual centres or trials are also of interest especially when there is considerable between-unit heterogeneity. However, failure to accommodate such heterogeneity if it exists can lead not only to an overprecise pooled estimate but also to incorrect inferences regarding unit level effects by not allowing these units to 'borrow strength' from similar units. Both issues have been addressed by the adoption of models similar in form to models (1) and (2) (Gray, 1994; Lambert and Abrams, 1995).

When between-unit heterogeneity exists, as with league tables, a reasonable question is to ask why? In these situations the exploration of the use of covariates to explain some of the between-unit variability has been advocated (Thompson, 1994). However, failure also to allow for the possibility of random variation can lead to an oversimplified interpretation of the between-unit differences. The adoption of *mixed effect* models in which both *fixed* covariate terms and *random* terms are included enables a more judicious assessment of both between-unit differences and the relative importance of unit level covariates (Breslow and Clayton, 1993).

The issue of model uncertainty raised in Section 6 is also important in both clinical trials and meta-

analysis. In the latter a dichotomy has developed between the adoption of the so-called *fixed effect* model and a model in which a random component is included, the *random effect* model. In the simplest case of discriminating between such models the usual test for heterogeneity has been shown to have low statistical power (Thompson, 1994), whereas in the mixed effect models the case is more complex. One approach that has been advocated in such situations is to obtain an overall estimate of effect which takes into account the uncertainty associated with the choice of model as well as the uncertainty associated with the estimates themselves (Abrams and Sansó, 1995).

The key message is that failure to model uncertainty correctly can lead to poor clinical or policy decisions being made at worst and unreliable comparisons at the very least.

Chris Chatfield (University of Bath): I welcome this paper as an important contribution to statistical issues in the calculation and presentation of league tables. The authors say (Section 6) that 'current official support for output league tables, even adjusted, is misplaced'. I think that they are perhaps being a little too polite and that this message needs to be stated more firmly and directly. One discussant for example appeared to be saying that, for all their acknowledged faults, current league tables give useful information and so no-one should want to suppress them. Well let me make it absolutely clear that (in some cases) I do. Some current league tables really are worse than useless. They are actually mischievous. When I hear of schools refusing to enter weaker pupils for the General Certificate of Secondary Education because they might ruin the school's statistics, I am greatly saddened. When I hear that schools are concentrating their efforts on middle range pupils to raise them from grade D to C, I despair. When I hear that teachers are encouraged to ignore the absences of some pupils, I can guess how reliable the resulting statistics on attendance are. The effect of a league table for New York surgeons was to make them rate their patients as being more seriously ill on presentation for treatment than they rated them before the league table was published. If league tables for surgeons become commonplace, I guess that a possible outcome could be that surgeons may eventually refuse to operate on those patients whose characteristics they judge to be such as potentially to spoil the surgeon's ranking. I could give further examples.

The authors say (Section 6) that the paper does not treat in detail 'the issue of quality and appropriateness of data in both adjustment and outcome measures'. However, I am sure that the authors would agree with me that these issues are just as important as the statistical modelling issues, and if some published league tables adversely affect both the quality and the appropriateness of the data on which they are based then they must expect to be disowned by statisticians.

Cindy L. Christiansen (Harvard Medical School, Boston): I commend the authors for their informative paper on issues and controversies surrounding league tables (also known as profiles). By discussing data, modelling and analyses, adjustments and the interpretation of results, they give an excellent summary of the complexities and statistical challenges involved in this work.

I want to emphasize four points related to health care profiling. First, the question 'what is quality?' is debatable; the answers to the question are powerful and affect other aspects of the health care system including access to care and costs. These effects should be purposeful and policy driven, not consequences with undesirable surprises.

Second, once quality has been defined, measuring and modelling variation in outcomes from non-controlled, non-experimental settings must be addressed. Thomas *et al.* (1993) discussed these problems by considering McAuliffe's (1984) dissection of the observed variability into three components: valid variance (the piece reflecting true performance differences), systematic variance and random error.

Our ability to estimate the valid component of the observed variance is controversial because it requires accurate estimates of the other two parts. To control for systematic variance across institutions, pertinent and complete data on all risks that affect outcome and that are not under the institution's control or related to quality are needed. Continued research on estimating systematic variance with available databases is crucial to profiling work. (Also see Rosen *et al.* (1995) and Greenfield *et al.* (1994).) Hierarchical models attempt to describe the random component of the observed variance and the 'valid', but unexplained, heterogeneity across institutions. As the authors have discussed, they are important statistical tools for profiling research. (Also see Morris and Christiansen (1995a) and Christiansen and Morris (1996)). However, I disagree with the comment in Section 3.1 that the choice of method is usually of philosophical rather than of practical importance. More research on method comparisons in profile analyses is needed before this position can be adopted.

Third, Bayesian models permit probability statements to be made that summarize performance and

rankings. In Morris and Christiansen (1995b) we suggested using probabilities to think about policy questions and programme assessments and to communicate results to decision makers, doctors and patients.

Finally, developing profile analyses must be collaborative work. A scientific profile analysis includes substantial input on appropriate case mix adjustments and on policy criteria from health care experts. Without this mixture of ideas and opinions, profiling research will fall short of its intended goals of measuring and promoting quality health care.

Russell Ecob (Medical Research Council Medical Sociology Unit, Glasgow): The issue of adjustment factors has been raised. This raises the question of which measures should ideally be included.

The authors' discussion of appropriate input measures, in the educational context, has cast some doubt on the validity of the sole use of an attainment measure at secondary school entry by drawing to our attention the issues of unreliability, possible effects of changing schools and the reported additional effect of the junior school attended (Goldstein and Sammons, 1996). It is well known that socio-economic factors have a continuing effect on progress in school, both junior (Mortimore *et al.*, 1988) and secondary (Fogelman, 1983; Garner and Raudenbush, 1991). League tables which neglect the socioeconomic composition of the schools are therefore penalizing those with disadvantaged social composition.

Returning to the reported junior school effect on secondary progress, could this again be an artefact of the lack of full socioeconomic controls? Junior schools generally have substantially smaller catchment areas than secondary schools and, given the area-based nature of social stratification, feeder schools to a given secondary school vary in their social composition. The junior school therefore acts as a proxy for these unmeasured social variables which are likely to operate both at the individual and the school level as a result of differential social composition of the neighbourhoods comprising the catchment areas of different junior schools (Garner and Raudenbush, 1991). On this reasoning, secondary schools taking their input from 'effective' junior schools would tend to have higher adjusted General Certificate of Secondary Education (GCSE) scores. Though effective junior schools may be expected to give pupils a lasting educational boost, an alternative possibility is that effective junior schools may be having transitory effects on attainment, elevating attainment above the pupil's hypothesized level of ability. This would result in their subsequent progress in secondary school and hence their adjusted GCSE score being reduced. Though both these processes probably operate, the former factors would lead to effects at junior school being positively correlated with effects at secondary school whereas the latter would lead to negative correlations.

Either way, achievement at 11 years (even if adjusted for reliability) can only partially measure the potential for GCSE attainment. Additional issues are raised when secondary schools select pupils on their initial attainment or on variables related to it. League tables controlling only for attainments, by leaving out other relevant factors, will always be imprecise and biased measures of school performance.

Antony Fielding (University of Birmingham): The central focus of the paper is uncertainty that is necessarily involved in any attempt to produce outcome indicators even when these involve proper adjustment within the context of a well-specified statistical model. The authors mention that official thinking appears to be moving in some ways towards the production of adjusted league tables in education. Yet the impression gained from a reading of some recent official reports is a preoccupation with single measures of institutional 'value added' however these are derived (see for example Schools Curriculum and Assessment Authority (SCAA) (1994) in addition to the authors' references). One problem with this, for example, is the important possibility that differential effectiveness of institutions might distort comparisons. The SCAA envisage finding individual residuals from a fitted national model and averaging within institutions. Thus, for instance, one institution which is as effective as another in terms of, say, intercept and the coefficient of the input variable may appear to have less value added if this latter coefficient is larger than the national value and the first institution has a much lower mean input. There are many problems similar to this if comparisons at different input levels are avoided in the search for simplicity of summary measures. My concern here is not that the official bodies are not aware of some of these or methods to address them. Rather it is that in the self-professed desire to avoid some 'technical niceties' any form of public transparency in the form of published league tables will avoid some central issues. This is, of course, additional to the fundamental problems of uncertainty raised by the authors.

There is an additional point about institutional comparisons briefly mentioned by the authors: the

central issue of the availability of resources and funding. In work on institutional-type differences in cost effectiveness (Thomas, 1990; Fielding, 1995) we find that comparisons of adjusted A-level outcomes are moderated when put into the context of unit costs. Class size of A-level groups is clearly a factor here but there are others related to different funding mechanisms governing a range of institutions. If current press reports are indicative this is also a concern being expressed by the Office for Standards in Education in reaction to the impending 1995 educational league tables.

Frank E. Harrell, Jr (University of Virginia, Charlottesville): This thoughtful paper could not be more timely. Hospitals and consortia of hospitals are frequently contracting with highly paid 'hired guns' to assist them in developing a (favourable) 'score-card'. Frequently such endeavours result in one of the following:

(a) a patient satisfaction survey in which virtually all patients happen to be very satisfied with the care that they received,
(b) an outcomes assessment in which there is inadequate adjustment for confounding (case mix) or
(c) an outcomes assessment in which an incorrect or improperly formulated statistical model was used.

These hired guns should be qualified by having them read and understand this paper and the papers that it cites.

The authors presented sage advice about risk adjustment and shrinkage and demonstrated the severe problems with ranking mortality rates. I would like to hear more from them about the adequacy of case mix adjustment. Even though adjustment for confounding ('treatment by indication') has been studied extensively in the statistical and epidemiological literature in the past 25 years, health outcomes analysts do not seem to have fully profited from this literature. For example, Hartz *et al.* (1992) used data sets collected for other purposes to determine the benefit of coronary angioplasty and coronary bypass surgery. The data sets excluded strong determinants of treatment selection, and in addition the authors chose to adjust only for 'significant' risk factors. Also, the duration of follow-up was inappropriately short. The same problems exist in comparing hospitals. Many analysts also seem to be confused in using a measure of explained variation in deciding whether the risk adjustment is adequate. Some feel that, if R^2 is sufficiently large, biases in comparing multiple hospitals have been removed. Often such analysts do not challenge this assumption by attempting to refine the base-line data collection and checking that expected mortality rates are relatively unchanged across hospitals. The opposite mistake is also being made, as witnessed by Green and Wintfeld's (1995) assertion that since an operative mortality model has a low R^2 it could not explain away real differences in case mix. The field could use detailed guidance on how to judge adequacy of adjustments. There may be a role for sensitivity analysis of the kind done in the propensity score context (Rosenbaum and Rubin, 1983).

Stephen Kendrick (National Health Service in Scotland, Edinburgh): The decision to publish clinical outcome indicators in Scotland was not taken lightly. The desire to promote a free flow of information relating to clinical outcomes throughout the Scottish Health Service was balanced against the understandable fear that the data would be used and interpreted in inappropriate ways which would create distress and lead to invalid decisions.

It was expected that one of the most common forms of misuse and misrepresentation would be to call the clinical outcome indicators 'league tables' and to use them to attempt to rank the quality of health care.

Thus great pains were taken at the press conference at which the indicators were released, in the text accompanying the measures and in all subsequent commentary to stress that the clinical outcome indicators are not league tables. They were not intended to be used as league tables, they were not presented as league tables and to an impressive extent, surprising to some, they have not been used as league tables within the Scottish Health Service. The mantra of the indicators is repeated again and again in the report:

> 'It is stressed that no direct inferences about quality of care should be drawn from these indicators. They are intended rather to highlight issues which may require further investigation'

(Clinical Resource and Audit Group, 1994). It could not be made clearer.

In the face of such efforts, two classes of commentator have been keenest to call the outcome

indicators league tables: an otherwise responsible Scottish popular press (e.g. 'Capital tops death league table') and some academic statisticians.

Most of the points made by Goldstein and Spiegelhalter about the limitations of the indicators are already set out in the report in which they were published. In general an even stronger line has been taken in Scotland. It has been emphasized that currently available administrative data, on their own, are an insufficient basis on which to prove or disprove the existence of differences in quality of care between hospitals (Kendrick *et al.*, 1995). The way forward in answering the questions posed by the indicators lies with more sensitive data and the involvement of the specialist clinicians concerned.

The involvement of statisticians in this enterprise is to be welcomed. However, from the perspective of the Scottish clinical outcome indicators this paper is largely a retreading of old ground and, by classifying them as league tables, misleading.

David Muxworthy (University of Edinburgh): The authors give relatively little attention to the effect of social environmental factors on school examination results. Home conditions for pupils may be expected to be related to input factors such as parental academic achievement and encouragement, facilities for homework and, indirectly owing to political pressure, to resources at the school itself.

An unpublished study that I undertook for a Scottish local authority in 1991 showed a very strong relationship between achievement in examinations and social deprivation. Taking for example, for each of the 51 secondary schools under the authority's control, the three measures

(a) the average number of Standard and Ordinary grade Scottish Certificate of Education 'passes' (grade 1–3) per pupil in the school year S4,
(b) the corresponding average for A–C marks at Higher grade in S5 and
(c) a simple measure of social deprivation, namely the percentage of free school meals taken at the school over all school years,

the correlation between (a) and (c) was -0.85 and between (b) and (c) was -0.79. These are graphically shown in Fig. 11.

With this information, the local authority could further investigate differences in school attainment, but the raw league tables as published invited erroneous conclusions to be drawn due to lack of relevant data.

Daphne Russell (University of Hull) **and Ian Russell** (University of York): This paper is important and timely, as the increasing proliferation of crude 'league tables' for institutional comparisons is causing widespread concern. Such a table may be misleading in three ways: the variables used may be inappropriate for comparing institutions, biased or imprecise. Although the authors address all three issues, the statistician's role in avoiding inappropriateness and bias needs greater emphasis. A precise

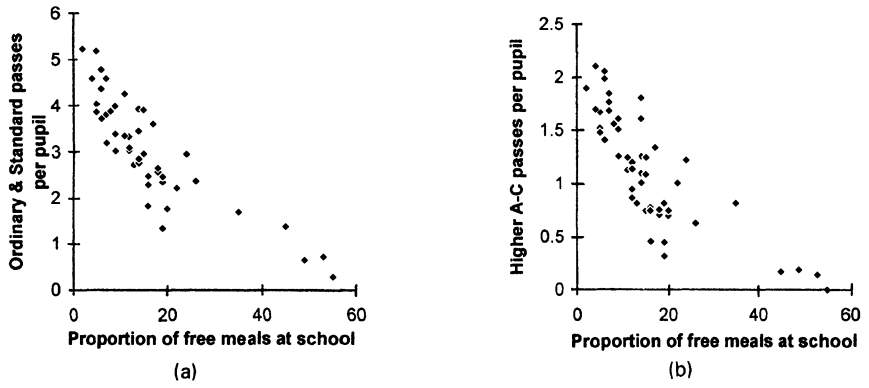


Fig. 11. (a) Plot of Ordinary and Standard grade passes per pupil against a deprivation index; (b) plot of Higher grade passes per pupil against a deprivation index

but inappropriate or biased measure is even more dangerous than an imprecise measure if it provides unjustified reassurance.

In the health field league tables of patient outcomes that make no attempt to adjust for case mix should be avoided; only adjusted tables have any claim to validity. This does not imply that the mere presence of adjustment (e.g. by a single risk score) is sufficient to solve the problem of bias. It is also essential that the method of adjustment be epidemiologically valid.

For example, the research literature contains many league tables comparing perinatal mortality between institutions or areas (despite its inability to measure morbidity in the child or mother). Most have standardized for differences in birth weight distributions, since these are effectively beyond the control of maternity services (e.g. Chalmers *et al.* (1978)). However, the traditional method of standardization is biased against areas with heavier births (Wilcox and Russell, 1983).

Wilcox and Russell (1986, 1990) later showed that a semiparametric approach can isolate without bias mortality differences that are independent of birth weight differences—the most appropriate for comparisons between institutions. Analysis of 1986 data for Scottish health boards showed that, although there were adverse birth weight distributions in three areas, only one also had a significant extra perinatal risk (Russell, 1995). That is the only difference that an unbiased, uncertainty conscious league table should have highlighted.

The authors' analysis of other Scottish data (Fig. 6) illustrates a potential weakness in the empirical Bayes approach. The populations of the island health boards are not only considerably smaller than those of the mainland boards but also sociologically very different. The assumption of prior ignorance is therefore unrealistic and leads to substantial and arguably biased changes in the position of island boards relative to large boards with less extreme rates. In such circumstances prior distributions should take full account of available local knowledge.

Finally, a rigorous evaluation is needed of the effect of league tables on outcomes, both included and excluded, and on the choices made by purchasers, providers and parents or patients.

Tom A. B. Snijders (University of Groningen): The uncertainty that is inherent in adjusted performance indicators mostly is large as well as practically important. In addition to uncertainty, there is another basic statistical *caveat*, which the authors do not mention explicitly: correlation does not imply causality. When output indicators are used for allocation of resources, the implicit assumption is made that the institutions are the cause of their performance level. Caution is also needed on this point.

The analysis hinges on the assumption of exchangeability of the residuals. This is plausible only if adequate adjustments have been made. Good adjustments will diminish our qualms about the assumption that the institutions are the cause of their level of performance. A sensible way of adjustment might also decrease the possibilities of manipulating performance discussed in Section 5.3. One example of a good approach to adjustment is the use, in the example in Section 4.2.1, of a bivariate analysis rather than an analysis with the first year's scores y_{1ij} as covariates. Initial scores are often used as covariates but in observational studies this usually introduces unwanted effects (this is known as 'Lord's paradox' in the psychometric literature; see Holland and Rubin (1983)).

The authors take a liberal view on the differences between Bayesian, likelihood and quasi-likelihood methods. With respect to estimation of uncertainty, I am hesitant to follow them. Given the frequent use of performance indicators, a frequency interpretation for the uncertainty intervals is desirable. The frequency interpretation of Bayesian intervals is not well established for 'difficult' parameters such as ranks. The standard errors produced by quasi-likelihood methods are not always reliable. Do the authors dare to make 19-to-1 bets on their Bayesian and quasi-likelihood 95% intervals?

For the analyses presented in Figs 6(b) and 7(b), it is striking as well as disturbing that the assumption of an exchangeable prior makes the ranks more uncertain than a model with board-specific fixed parameters. Ranks have, of course, an inherent mutual dependence. The approach taken in Section 4, where the focus is on pairwise comparisons, seems more relevant than the approach in Section 5 where 95% intervals are given for single rates and ranks. In a Bayesian analysis of pairwise differences of rates or ranks of health boards, does the exchangeable prior still lead to more uncertainty than board-specific parameters? I would like to invite the authors to develop a further analysis and an interpretation of the effects that models with exchangeable prior distributions for level 2 parameters have on the posterior distribution of the ranks.

Neil H. Spencer (Staffordshire University, Stoke-on-Trent): To my mind, two issues are emphasized in this paper. The first concerns the model or measure of performance used. This must be chosen so that

the results obtained are relevant to their audience, and I suggest that it may be necessary on occasions to construct more than one performance indicator for each institution to reflect the specific interests of different bodies of people. The second issue is that of interpretation, and it is to be hoped that the media and public will, over the coming years, become more aware of the uncertainty that should be attached to the performance indicators. Perhaps if, as may be expected, future league tables show notable changes in the rankings of institutions from year to year then this awareness will develop in relatively little time.

I would contend that there is a third issue: that of actually obtaining valid performance indicators from the model chosen. For example, a simple multilevel model to obtain performance indicators for schools on the basis of A-level results, taking into account General Certificate of Secondary Education (GCSE) results as an explanatory variable, suffers from having a random effect for each school in the random part of the model that will not be independent of the GCSE points score if the GCSEs and A-levels are taught in the same institution (see Spencer and Davies (1995a) for more details). This may lead to inconsistent parameter estimates being obtained for the model and unreliable performance indicators may result. Consequences for the ranking of schools may be considerable (Spencer and Davies, 1995b). Solutions to this problem when more than two measures of attainment per pupil are available are to be found in Spencer and Davies (1995a, b), and my on-going research provides a solution when only two measures per pupil are available, as is the case when studying GCSE and A-level results.

It is to be welcomed that the Society is taking an active interest in this controversial subject, and the authors of this paper are to be congratulated for presenting the issues surrounding it in a very even-handed manner. However, care must always be taken to ensure that performance indicators used are valid otherwise any conclusions drawn may be unreliable.

William Tarnow-Mordi (University of Dundee) **and Gareth Parry** (University of Sheffield):

Inappropriate and appropriate comparisons of intensive care units using the clinical risk index for babies and paediatric risk of mortality scores

The paper which Professor Goldstein and Dr Spiegelhalter quoted by de Courcy-Wheeler *et al.* (1995) exemplifies some of the problems in comparing risk-adjusted mortality between institutions. In a cohort of 643 infants treated in 18 hospitals in one year, these investigators compared mortality in larger *versus* smaller neonatal intensive care units, using the clinical risk index for babies (CRIB) to adjust for differences in case mix. The difference in risk-adjusted mortality between larger and smaller units was not statistically significant. They erroneously concluded that the larger hospitals had excess mortality, although the 95% confidence intervals around the point estimates for risk-adjusted mortality in the two groups were wide and overlapping. In our original paper describing the CRIB score (International Neonatal Network, 1993) we constructed 'league tables' based on admissions to individual hospitals over 2 years. These showed that, even with 2 years' data, the numbers were too small to allow fine ranking of individual hospitals. By the time that samples sufficiently large for more reliable ranking can accumulate, the results may be irrelevant.

A more appropriate strategy was illustrated in a national, prospective, stratified random sample of 16 US paediatric intensive care units using the paediatric risk of mortality score to adjust for case mix in a cohort of 5415 children (Pollack *et al.*, 1994). Testing prespecified hypotheses, the risk-adjusted mortality was about 80% greater in teaching hospitals (relative odds of dying 1.79; 95% confidence interval (CI) 1.23–2.61), and about 30% lower in hospitals with a paediatric intensive care specialist (relative odds 0.65; 95% CI 0.44–0.95). Scoring systems seem most appropriate for investigating risk-adjusted mortality in relation to organizational characteristics of groups of hospitals, if these can yield sufficiently large samples for adequate power in a reasonable time.

The **authors** replied later, in writing, as follows.

We are grateful to all the discussants for supporting our view that the issue of measuring institutional performance raises important problems and that the Royal Statistical Society (RSS) has a useful role to play. We shall deal first with the main substantive issues, then with the technical ones and finally outline further areas for study and action. However, in the available space we cannot hope to answer all the concerns raised by the discussants, and we apologize for questions left unanswered.

There is general agreement among discussants that uncontextualized league tables based on performance measures are undesirable and often very misleading. Many of the discussants make positive recommendations for improving the measurement and reporting of information summarizing the context, and we shall comment on some of these below.

Rosemary Butler is the only discussant who seems to take the view that the justification for the publication of existing performance tables in health is that 'they have been effective' and that their publication 'leads to improved performance by stimulating informed questioning'. One difficulty for her position is that the demonstration of 'effectiveness' is extremely problematic and we know of no experimental attempt to carry out such evaluations in health or education, save for the interesting, but still in progress, experiment described by Dr Brand. Observational evaluation studies are fraught with difficulties: although it has been reported that the risk-adjusted mortality for cardiac artery bypass grafts in New York State dropped from 4.2% to 2.7% since the introduction of the 'report cards', 66% of this change was attributable to the increased reported severity in patients, and it is difficult to separate the remainder from background improvements throughout the USA (Green and Wintfeld, 1995).

Butler also appears to feel that messages on uncertainty should not be given since they may be misinterpreted by the popular press. It seems unfortunate, to say the least, for dissemination policy to be dictated by the lowest common denominator, although we suspect that those who set the 'business agenda' may also have problems with any expression of uncertainty (see Bartholomew's contribution). Fortunately not all health agencies take this view: a recent report comparing *in vitro* fertilization clinics (Health Fertilisation and Embryology Authority, 1995) did give unadjusted live-birth rates, but then in larger print provided adjusted live-birth rates, with uncertainty intervals. Our view, and that of many other discussants, is that it is one of the important roles of our Society to undertake to educate the public and the media on the statistical issues. In discussing the public dissemination of comparative information, the Chief Social Affairs Correspondent of the British Broadcasting Corporation has argued that 'differences should be explored and explained', and that 'it should not be beyond the intelligence of the rest of us to understand the limitations of this information' (Dickson, 1995). Rosemary Butler appropriately chastises us for overlooking the definition of a 'significant' change given in the current National Health Service performance tables (National Health Service Executive, 1995), but this definition, in terms of a 5% change, only serves our point. Since denominators are not provided we cannot work out whether such an observed decrease may be due to chance alone, thus unfairly labelling an institution as deteriorating. Over 100 years ago, a classic paper to this Society (Edgeworth, 1885) introduced the proper allowance for uncertainty, and it is surely time for this message to be heeded.

Carol Fitz-Gibbon also refers to the 'success' of feed-back of performance indicator information about immunization rates. Unfortunately, as Goodhardt's example shows, the mere observation of subsequent 'improvements' is poor evidence for adducing causation. Fitz-Gibbon refers to the way in which institutions will 'play the system' in ways which may be harmful, and this encouragement to 'gaming' is also well illustrated in contributions from Nick Longford, Eastwood and her colleagues, Martin McKee and Chris Chatfield. Fitz-Gibbon makes the useful point that many performance measures in education are too highly aggregated and that more precisely focused measures are desirable. The problem, however, is that in reporting, say, science performance rather than all-subject averages we are typically dealing with very much smaller populations of students with resulting increases in uncertainty intervals. A similar problem exists in health with the reporting of individual specialities. As with the examples that we quote in the paper, when the uncertainty becomes so large that most institutions cannot be distinguished, there is a strong case for reporting just this: namely that the performance indicator in question conveys little useful discriminatory information. Carol Fitz-Gibbon, as well as David Draper, Tom Snijders and other discussants, raises the question about causal inference and randomized experiments. Our discussion of the need to adjust or contextualize performance measures recognizes that, within a non-experimental framework, we need to seek out those factors which theoretical or other considerations suggest are both influential and relevant; hence our discussion of type A and type B effects. We agree with Draper that not all influential factors are necessarily relevant, but our example in Section 3.1 of localities and general practitioners is in our view relevant when comparing hospitals. We must disagree with Fitz-Gibbon's advocacy of ordinary least squares over multilevel models. The complexity of the systems being studied requires the deployment of correspondingly complex models, and a good discussion of this issue in the area of school performance indicators can be found in Aitkin and Longford (1986).

John Gardner reinforces the distinction between type A and type B inferences and he makes the important point that parents and students, and presumably patients, want to know what the expected outcome will be in different kinds of institutions given their own characteristics. This implies both that we need to make proper adjustments for inputs and also to take note of the possibility of differential institutional effectiveness as we illustrate in Fig. 1. For both these reasons raw tables are misleading and

it is difficult to see why anyone who is aware of these issues would really want to use them to judge an institution. Gardner worryingly quotes a figure of half of a sample of parents claiming that existing league tables are useful in choosing a school. It would be an interesting experiment to provide a randomly chosen group of parents with information about the deficiencies of league tables and to study how their behaviour and beliefs changed as compared with a non-informed control group. We shall return to the issue of public education in our final section.

Sheila Gore provides an important example of where the use of performance measures attributed to individual institutions is a clearly inferior way of obtaining the required information. This is both because it encourages less than accurate measurement and because a properly conducted trial is scientifically preferable. This point applies more generally in health and education. From a long-term point of view it is knowledge of what works and why which will lead to improvements; the ranking of institutions *per se* will not provide that information unless it enables us to approach causal inferences about influential factors. In this sense, concentrating resources on naïve institutional comparisons distracts both money and people from the wider and far more important task.

We were delighted to be informed by Alison Macfarlane that Florence Nightingale had made our main points well over a century ago! Macfarlane raises the issue of how the requirement to report data can subtly influence their collection, and by implication that this is a problem which needs to be studied carefully. She also discusses the difficulty of classifying measures into the categories of input, process and output which again raises the issue of distinguishing between type A and B effects. She makes the point also made by David Bartholomew that coarse groupings will be more appropriate than fine rankings. As Bartholomew points out, the group boundaries should be determined by our knowledge of the uncertainties associated with each institution and a useful construction principle would be to minimize the overall (weighted) probability of 'incorrectly' ranking each pair of institutions. A similar principle underlies our 'overlap' intervals and some further work on this would be useful.

Sharon-Lise Normand points to the difficulties associated with using unstandardized measurements such as those based on physicians' notes. This is a continuing concern where performance systems are forced to rely on information that happens to be available because of the expense of collecting properly structured and comparable data. She also emphasizes the important distinction between choosing covariates to improve predictive power and choosing those which are truly endogenous to the system. We particularly agree with her that institutions need to be concerned with those influential factors over which they have some control, and this again emphasizes the need for consideration of type B analyses, a point also made by Ramesh Kapadia.

The analysis of hospital processes described by David Draper is an extremely interesting model for evaluating the usefulness of adjusted league tables. Considerable difficulties are of course associated with obtaining good process measures, but we would certainly commend studies of the kind proposed by Draper to any government agency proposing to introduce adjusted performance tables. David Draper's remarks on time homogeneity are well taken and in our view all attempts at performance indicator construction should be run over several years to study the stability of predictions. In education existing research suggests that there is considerable year-to-year variation which requires extreme caution in interpreting the result from a single cohort (Gray *et al.*, 1995).

Mary Smalls and Steve Kendrick query why the Scottish data were used, when their reports are full of warnings not to make naïve judgments about reasons for observed differences, and they carefully avoid all temptation to rank. First, we apologize for not emphasizing their *caveats* in our paper and presentation: we acknowledge the care and effort put into the Scottish tables, and in particular admire the 15 years of effort to get intervals included (and we hope that it does not take a similar period south of the border!). Their data were used precisely to illustrate the dangers of naïve ranking without fully acknowledging uncertainty. However, we do not accept that our attempts to model institutional differences are simply adding 'to the complexity'. The complexity is there, whether or not management scientists, accountants or politicians are prepared to acknowledge it; our job is to model it as accurately as possible rather than to provide an oversimple description. Stephen Kendrick admits that sections of the media used the Scottish data to produce league tables even where the official publication issued warnings. In our view it is insufficient simply to wring one's hands over this. A major purpose of our paper is to open a discussion of ways in which such misuses can be minimized by appropriate presentation of uncertainty.

We have no real disagreements with the issues set out for consideration by John Copas and we share his frustration at statisticians' work being used out of context. There is certainly something here which the RSS should consider.

Ramesh Kapadia makes the important point about educational league tables that the choice of scoring system can be important. We do not follow his implication that because 'output indicators' will be used for resource allocation then we need to reserve our criticism. The point, surely, is that the inadequacies of these indicators, as we point out in our conclusions, suggests caution over using them for resource allocation.

Ian Schagen, like Carol Fitz-Gibbon, argues for the identification of institutions or departments with 'extreme' values and then studying their characteristics. As part of an exercise where such institutions are compared with 'non-extreme' institutions this seems to have merit. Ian Diamond and Fiona Steele give a good example where the study of non-extreme as well as extreme cases yields useful information. In their case there was plenty of locally relevant information which could be used to inform judgments. We agree with attempts to provide information on multidimensional outcomes, whether at departmental level and in affective as well as cognitive domains in schools or for the several possible measures of performance associated with hospitals as pointed out by Neil Spencer.

Alastair Leyland also makes an important point about modelling institutional differences within the full complexity of coexisting geographical and social structures. The importance of introducing crossing factors into purely hierarchical data structures is also mentioned in Section 3.1 for educational data, and this should inform further work. A similar point is made by Neil Spencer who suggests that knowledge of the institution where the General Certificate of Secondary Education is taken is relevant to modelling A-level institutional differences.

Stephen Senn refers to school size and class size as potentially influential factors. The effect of school size is generally not found to be large: class size, at least in part, is a process variable and there is a considerable debate about its importance. The issue of selecting students on the basis of A-level results is quite separate from comparing schools. As a selection device for individuals, an A-level result is an attempt to summarize achievement and although it may be influenced by the school attended there is no *a priori* reason for taking further account of school factors, unless we believe that, for example, such factors may add predictive power in terms of university performance. The shrinkage issue is irrelevant here.

As we mentioned in the case of examination results William Tarnow-Mordi and Gareth Parry emphasize the problem of institutional comparisons in health becoming of historical relevance only, by the time that sufficient numbers have been accumulated. It seems to us that this issue alone will often destroy the usefulness of institutional comparisons.

Turning to the more technical contributions, several discussants commented on the use of shrinkage estimators. As David Draper points out, the best justification for this comes from an underlying concern with prediction for an institution, and as with all predictions we will produce a (shrunken) estimate together with an estimate of its uncertainty. Posterior shrunken means in this sense are no different from ordinary regression predictions: if we wish to predict a value for a new institution that is not represented in our analysis, we will likewise obtain a shrunken estimate conditional on the observed values of the variables measured for the individuals in that institution. Institutions where there are few individuals will generally provide poor predictors for that institution's underlying parameter values, which makes us tend to agree with Rex Galbraith that this supports the case against ranking rather than making a case against shrinking. We also agree with Rex Galbraith that, where possible, it will be preferable to provide predictions based on the analysis of large standardizing samples, although we do need to be concerned about stability over time. Nick Longford also makes this point about the analysis of small numbers of schools within one education authority. Goodhardt makes an excellent practical case for hierarchical modelling, and Abrams and Lambert point out that similar concerns are leading to such models being explored in many other areas. Table 4 and Fig. 9 presented by Galbraith are interesting and we accept that there is much scope for further work on presentation.

We are grateful to Tom Louis and Nick Longford for pointing out the distinction between ranking posterior means and estimating expected posterior ranks, and we apologize for overlooking Louis's work with Laird. In the education examples we are in fact estimating means rather than ranks, and which presentation is preferable is an interesting question in any given application.

Cindy Christiansen criticizes us for downplaying the differences between competing methods of statistical inference. We do not deny that these differences are important: we merely wish to emphasize that whichever approach is used yields similar kinds of conclusions about the role of institutional comparisons. This is well illustrated by Ian Langford and Alastair Leyland's analysis of the surgery data using the bootstrap as opposed to Gibbs sampling. We (the authors) adopt different statistical philosophies, but we are in agreement that it would be unfortunate if the discussion between statisticians were to be side-tracked into a debate on this issue.

We agree with Frank Harrell, Keith Abrams and Paul Lambert that the sensitivity of the model to particular choices and assumptions is important both within a formal framework and with respect to a choice of covariates and so on, and an interesting example where model choice makes a difference is supplied by David Draper. A fully Bayesian approach would allow a posterior probability of competing models to be calculated, thus answering John Copas's and Jane Galbraith's concern for full expressions of uncertainty in the model specification, although this might be becoming a little too complex. Jane Galbraith is also concerned with allowing for the uncertainty in variance estimates. When using a fully Bayesian approach we obtain consistent interval estimators by virtue of carrying through the full uncertainty in the estimation and, as Longford points out, can easily include prior information to increase stability. With the generalized least squares (maximum likelihood) approach we can use either delta method adjustments or bootstrapping as done by Ian Langford and Alastair Leyland to provide consistent estimators. This also addresses the concern of Tom Snijders about standard errors for quasi-likelihood methods. In the paper in Section 3.1 in fact, 'plug-in' estimates were used which are conservative and tend somewhat to underestimate the interval widths. Nick Longford has picked this up, although we would point out that Figs 2 and 3 are designed to control the overall type 1 error rate only for pairwise comparisons.

In contrast with the strong recommendation for the Bayesian approach by Louis and Christiansen, Tom Snijders would welcome proven frequentist properties for the Bayesian interval estimates on the ranks. He is right that these properties have not been explored, and we intend, at least, to check the empirical coverage for predictions based on these methods. With regard to pairwise comparison of ranks, it is easy from the Gibbs sampling output to estimate the probability that any particular pair of institutions is ranked in a particular order: these estimates can be attractively displayed as shaded areas on a grid. Assuming exchangeable priors tends to increase the overlap and hence the probability for any particular order tends to 0.5 and so the grid becomes more uniformly shaded.

Finally we would like to make some general remarks about the future of institutional comparisons.

It seems important, as Martin McKee and others have pointed out, that there is a wider public and professional discussion of the issues raised in the paper and by the discussants. We believe that there are three key issues: adjustment, uncertainty and the multidimensionality of outcomes. We think that these can be understood widely if presented with care to those who are willing to listen. We have been encouraged by the example of education where, in the UK over the past 5 years, the profession, important sections of the public and even the government itself have accepted in principle the case for adjustment (value added). An important event here, we believe, was the publication by *The Guardian* newspaper of a supplement devoted to a presentation of adjusted A-level results together with uncertainty intervals. To educate people to recognize uncertainty is, we suspect, a more difficult task and we would urge the Society to give careful thought to ways in which this could be tackled. Chris Chatfield makes the point that statisticians should also be concerned about the issue of data quality and we agree whole-heartedly.

One clear way in which these issues can be brought home to people is to carry out sound evaluations of the effects of the publication of institutional performance data. Carol Fitz-Gibbon and Gerald Goodhardt call for experiments, and such studies would expose the harmful effects associated, for example, with gaming, and would be able to point out the need for adequate adjustments and allowance for uncertainty. We are encouraged by the Dutch example described by Dr Brand.

It seems to us important to recognize, as Nick Longford points out, that even if we could achieve accurate institutional comparisons these would not of themselves necessarily tell us how we should allocate resources. Our paper is only concerned with the very first step in this process: that of fairly identifying contextualized differences between institutions. The vital questions are then to understand why institutions differ and what action is likely to bring about improvements: these are certainly areas where statistical expertise has much to offer.

We are most grateful to the Society for providing the opportunity to debate these issues, and we hope that it will provide a basis for the Society's continuing involvement in this difficult but important area.

REFERENCES IN THE DISCUSSION

Abrams, K. R. and Sansó, B. (1995) Model estimation and discrimination in meta-analysis—a Bayesian perspective. *Technical Report 95-03*. Department of Epidemiology and Public Health, University of Leicester, Leicester.

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies (with discussion). *J. R. Statist. Soc.* A, **149**, 1–42.

American College of Cardiology and American Hospital Association (1990) Guidelines for the early management of patients with acute myocardial infarction. *J. Am. Coll. Card.*, **6**, 249–292.

Audit Commission (1995) *For Your Information: a Study of Information Management and Systems in the Acute Hospital.* London: Her Majesty's Stationery Office.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Burdett, H. C. (1882) The relative mortality, after amputations, of large and small hospitals, and the influence of the antiseptic (Listerian) system upon such mortality (with discussion). *J. Statist. Soc. Lond.*, **45**, 444–483.

Chalmers, I., Newcombe, R. and West, R. (1978) Adjusted perinatal mortality rates in administrative areas of England and Wales. *Hlth Trends*, **10**, 24–29.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference (with discussion). *J. R. Statist. Soc.* A, **158**, 419–444.

Christiansen, C. L. and Morris, C. N. (1996) Fitting and checking a two-level Poisson model: modeling patient mortality rates in heart transplant patients. In *Bayesian Biostatistics* (eds D. Berry and D. Stangl), pp. 467–501. New York: Dekker.

Clarke, A. and McKee, M. (1992) The consultant episode: an unhelpful measure. *Br. Med. J.*, **305**, 1307–1308.

Clinical Resource and Audit Group (1994) *Clinical Outcome Indicators: December 1994.* Edinburgh: Scottish Office.

Colombo, R. A. and Morrison, D. G. (1988) Blacklisting social science departments with poor PhD submission rates. *Mangmnt Sci.*, **34**, 696–706.

de Courcy-Wheeler, R. H. B., Wolfe, C. D. A., Fitzgerald, A., Spencer, M., Goodman, J. D. S. and Gamsu, H. R. (1995) Use of the CRIB (clinical risk index for babies) score in prediction of neonatal mortality and morbidity. *Arch. Dis. Childhd*, **73**, F32–F36.

Department of Health (1993) *Population Health Outcome Indicators for the NHS: a Consultation Document.* London: Department of Health.

Devine, O. J., Louis, T. A. and Halloran, M. E. (1994) Empirical Bayes estimators for spatially correlated incidence rates. *Environmetrics*, **5**, 381–398.

Dickson, N. (1995) League tables: use for patients. *Qual. Hlth Care*, **4**, 1.

Draper, D. (1995) Inference and hierarchical modeling in the social sciences (with discussion). *J. Educ. Behav. Statist.*, **20**, 115–147, 228–233.

Edgeworth, F. Y. (1885) Methods of statistics. *J. Statist. Soc. Lond.*, Jubilee vol., 181–217.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap.* London: Chapman and Hall.

Fielding, A. (1995) Institutional disparities in the cost-effectiveness of GCE A level provision: a multilevel approach. *Educ. Econ.*, **3**, 159–172.

Fogelman, K. (ed.) (1983) *Growing up in Great Britain: Papers from the National Child Development Study.* London: Macmillan.

Galbraith, R. F. (1988) Graphical display of estimates having differing standard errors. *Technometrics*, **30**, 271–281.

——— (1994) Some applications of radial plots. *J. Am. Statist. Ass.*, **89**, 1232–1242.

Garner, C. L. and Raudenbush, S. W. (1991) Neighborhood effects on educational attainment: a multilevel analysis. *Sociol. Educ.*, **64**, 251–262.

Ghosh, M. (1992) Constrained Bayes estimates with applications. *J. Am. Statist. Ass.*, **87**, 533–540.

Goldstein, H. (1995a) *Multilevel Statistical Models.* London: Arnold.

——— (1995b) *Delta Method Adjustments to Covariance Matrix Estimators for Random Parameters and Residuals.* London: Institute of Education.

Goldstein, H. and Sammons, P. (1996) The influence of secondary and junior schools on sixteen year examination performance. In *School Effectiveness and School Improvement.* To be published.

Gray, J., Jesson, D. and Goldstein, H. (1995) *Changes in GCSE Examination Performance using Value Added Analysis over a Five Year Period in One Local Education Authority.* Cambridge: Homerton College.

Gray, J. and Wilcox, B. (1995) *Good School, Bad School: Evaluating Performance and Encouraging Improvement.* Buckingham: Open University Press.

Gray, R. J. (1994) A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, **50**, 244–253.

Green, J. and Wintfeld, N. (1995) Report cards on cardiac surgeons: assessing New York State's approach. *New Engl. J. Med.*, **332**, 1229–1232.

Greenfield, S., Sullivan, L., Silliman, R. A., Dukes, K. and Kaplan, S. H. (1994) Principles and practice of case mix adjustment: applications to end-stage renal disease. *Am. J. Kid. Dis.*, **24**, 298–307.

Guy, W. A. (1856) On the nature and extent of the benefits conferred by hospitals on the working classes and the poor. *J. Statist. Soc. Lond.*, **19**, 12–27.

Harris, A., Jamieson, I. and Russ, J. (1995) A study of 'effective' departments in secondary schools. *School Organisn*, **15**, 283–299.

Hartz, A. J., Kuhn, E. M., Pryor, D. B., Krakauer, H., Young, M., Heudebert, G. and Rimm, A. A. (1992) Mortality after coronary angioplasty and coronary artery bypass surgery (the national Medicare experience). *Am. J. Card.*, **70**, 179–185.

Health Fertilisation and Embryology Authority (1995) *The Patients' Guide to DI and IVF Clinics*. London: Health Fertilisation and Embryology Authority.
Holland, P. W. and Rubin, D. B. (1983) On Lord's paradox. In *Principles of Modern Psychological Measurement* (eds H. Wainer and S. Messick), pp. 3–35. Hillsdale: Erlbaum.
Iezzoni, L. (ed.) (1994) *Risk Adjustment for Measuring Health Care Outcomes*. Ann Arbor: Health Administration Press.
International Neonatal Network (1993) The CRIB (clinical risk for babies) score: a tool for assessing initial neonatal risk of death and comparing the performance of neonatal intensive care units. *Lancet*, **342**, 193–198.
Jenkins, S. (1995) *Accountable to None: the Tory Nationalisation of Britain*. London: Hamilton.
Kahn, K., Brook, R., Draper, D., Keeler, E., Rubenstein, L., Rogers, W. and Kosecoff, J. (1988) Interpreting hospital mortality data: how can we proceed? *J. Am. Med. Ass.*, **260**, 3625–3628.
Kahn, K., Rubenstein, L., Draper, D., Kosecoff, J., Rogers, W., Keeler, E. and Brook, R. (1990) The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: an introduction to the series. *J. Am. Med. Ass.*, **264**, 1953–1955.
Kendrick, S., Barnwell, E. and Boyd, J. (1995) Clinical outcome indicators in Scotland: the first steps. *A. Conf. European Health Management Association, Celle, July*.
Kendrick, S. and Clarke, J. (1993) The Scottish record linkage system. *Hlth Bull.*, **51**, 72–79.
Kerr, S. (1975) On the folly of rewarding A while hoping for B. *Acad. Mangmnt J.*, **18**, 769–783.
Kornai, J. (1992) *The Socialist System: the Political Economy of Communism*. Oxford: Clarendon.
Laird, N. M. and Louis, T. A. (1989) Empirical Bayes ranking methods. *J. Educ. Statist.*, **14**, 29–46.
Lambert, P. C. and Abrams, K. R. (1995) Meta-analysis using multilevel models. *Multilev. Modllng Newslett.*, **7**, no. 2, 17–19.
Lancet (1995) Leap of faith over the data tap. *Lancet*, **345**, 1449–1451.
Louis, T. A. (1984) Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Statist. Ass.*, **79**, 393–398.
McAuliffe, W. E. (1984) A validation theory for quality assessment. In *Hospital Quality Assurance: Risk Management and Program Evaluation* (eds J. J. Pena *et al.*). Germantown: Aspen.
McColl, A. J. and Gulliford, M. C. (1993) *Population Health Outcome Indicators for the NHS: a Feasibility Study*. London: Royal College of Physicians.
Morris, C. N. and Christiansen, C. L. (1995a) Fitting Weibull duration models with random effects. *Lifetime Data Anal.*, **1**, 347–359.
——— (1995b) Hierarchical models for ranking and for identifying extremes. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.
Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988) *School Matters; the Junior Years*. London: Chapman.
National Health Service (1995) *The Patients' Charter: the NHS Performance Guide 1994–95*. London: Department of Health.
National Health Service Executive (1995) *The NHS Performance Guide 1994–1995*. Leeds: National Health Service Executive.
Nightingale, F. (1863) *Notes on Hospitals*, 3rd edn. London: Longman, Green, Longman, Roberts and Green.
Normand, S. L., Glickman, M. E. and Ryan, T. (1995) Modeling mortality rates for elderly heart attack patients: profiling hospitals in the Cooperative Cardiovascular Project. In *Case Studies in Bayesian Statistics* (eds C. Gatsonis, J. Hodges, R. Kass and N. Singpurwalla). New York: Springer.
Persaud, R. (1995) Inability to reason statistically is prime cause of lottery fever. *Br. Med. J.*, **311**, 1225.
Pollack, M. M., Cuerdon, T. T., Patel, K. M., Ruttimann, U. E., Getson, P. R. and Levetown, M. (1994) Impact of quality of care factors on pediatric intensive care mortality. *J. Am. Med. Ass.*, **272**, 941–946.
Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*. London: Institute of Education.
Robinson, G. K. (1991) That BLUP is a good thing: the estimation of random effects. *Statist. Sci.*, **6**, 15–51.
Rockall, T. A., Logan, R. F., Devlin, H. B. and Northfield, T. C. (1995) Variation in outcome after acute upper gastrointestinal haemorrhage. *Lancet*, **346**, 346–350.
Rosen, A. K., Ash, A. S., McNiff, K. J. and Moskowitz, M. A. (1995) The importance of severity of illness adjustment in predicting adverse outcomes in the Medicare population. *Clin. Epidem.*, **48**, 631–643.
Rosenbaum, P. R. and Rubin, D. B. (1983) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B, **45**, 212–218.
Russell, D. (1995) Birthweight and perinatal mortality: development and application of the Wilcox–Russell model. *PhD Thesis*. University of Aberdeen, Aberdeen.
Sackett, D. L., Haynes, R. B., Guyatt, G. H. and Tugwell, P. (1991) *Clinical Epidemiology*, 2nd edn. Boston: Little, Brown and Co.
Schools Curriculum and Assessment Authority (1994) *Value Added Performance Indicators for Schools*. London: Schools Curriculum and Assessment Authority.
Scottish Office (1994) *Clinical Outcome Indicators—1993*. Edinburgh: Clinical Resource and Audit Group.
Smith, P. (1990) The use of performance indicators in the public sector. *J. R. Statist. Soc.* A, **153**, 53–72.
——— (1995) On the unintended consequences of publishing performance data in the public sector. *Int. J. Publ. Admin.*, **18**, 277–310.

———(ed.) (1996) *Outcome Measurement in the Public Sector*. London: Taylor and Francis.
Spencer, N. H. and Davies, R. B. (1995a) Consistent parameter estimation for lagged multilevel models. To be published.
———(1995b) Using multilevel models to create value-added league tables for schools. To be published.
Statistical Society of London (1842) Report of the Committee on hospital statistics. *J. Statist. Soc. Lond.*, **5**, 168–176.
———(1844) Second report of the Committee of the Statistical Society of London on hospital statistics. *J. Statist. Soc. Lond.*, **7**, 214–231.
Steele, F., Diamond, I. and Amin, S. (1996) Immunization uptake in rural Bangladesh: a multilevel analysis. *J. R. Statist. Soc.* A, **159**, 289–299.
Steele, J. C. (1861) Numerical analysis of the patients treated in Guy's Hospital in the last seven years, from 1854 to 1861. *J. Statist. Soc. Lond.*, **24**, 374–401.
———(1877) The mortality of hospitals, general and special in the United Kingdom, in times past and present. *J. Statist. Soc. Lond.*, **40**, 177–261.
Thomas, H. (1990) *Education Costs and Performance: a Cost-effectiveness Analysis*. London: Cassell.
Thomas, J. W., Holloway, J. J. and Guire, K. E. (1993) Validating risk-adjusted mortality as an indicator for quality of care. *Inquiry*, **30**, 6–22.
Thompson, S. G. (1994) Why sources of heterogeneity in meta-analysis should be investigated. *Br. Med. J.*, **309**, 1351–1355.
Tukey, J. W. (1974) Named and faceless values: an initial exploration in memory of Prasanta C. Mahalanobis. *Sankhya* A, **36**, 125–176.
Wilcox, A. J. and Russell, I. T. (1983) Perinatal mortality: standardising for birthweight is biased. *Am. J. Epidem.*, **118**, 857–864.
———(1986) Birthweight and perinatal mortality: III, Towards a new method of analysis. *Int. J. Epidem.*, **15**, 188–195.
———(1990) Why small black infants have a lower mortality rate than small white infants: the case for population-specific standards for birth weight. *J. Pediatr.*, **116**, 7–10.