League tables and their limitations: statistical issues in comparisons of institutional performance.

by

Harvey Goldstein
Institute of Education
20 Bedford Way
London, WC1H 0AL, U.K.

and

David J. Spiegelhalter
MRC Biostatistics Unit
Institute of Public Health
Robinson Way
Cambridge, CB2 2SR, U.K.

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 institutional comparisons 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 baseline measures and the need to exercise care and sensitivity when interpreting apparent differences. The paper emphasises that statistical methods exist which can contribute to an understanding of the extent and possible reasons for institutional differences. It also urges caution by discussing the limitations of such methods.

Keywords

Multilevel modelling, physician profiling, ranking, risk shrinkage estimators, stratification, value added, Markov chain Monte Carlo.

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.

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 (NHS Executive, 1995) and local government services (Audit Commission, 1995) have had attention. Smith (1990) discusses the background to this interest and looks at the social, economic and political purposes performed by performance indicators in both the private and 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 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 organisational structure, while '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: both within 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 considerations as well as social and political ones 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 do 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 emphasising a statistical perspective that is common to any subject-matter area, we shall argue for the proper contextualisation 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. While there are many technicalities 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 describe a basic framework for some issues that are common to all attempts to compare institutions 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 using practical examples. While 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 summarises 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 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 a quite different measurement of achievement, ability etc. Likewise, for measurements such as surgical success rates, we would wish to contextualise these using measures such as severity of 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 ones where institutions will be judged against each other, conditional on prior or baseline measurements, and so the indicators we shall refer to in this paper are 'comparative'. In some cases, for example immunisations, it may be possible to achieve agreement on 'absolute' standards of performance. While similar statistical issues will arise in those cases, we shall not be concerned with them. The typical, if somewhat oversimplified, case is exemplified in Figure 1 which shows a comparison between two institutions in terms of predicted mean output for given input or baseline measure. Institution A has a higher expected achievement than B for individuals with low baseline scores and vice versa for those with high baseline scores. The issue therefore, is the comparative one 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’.

(Figure 1 here)

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 student 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 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 organisation or teaching styles and we shall have more to say about such variables later in both health and education.

While the focus of the present paper is on the use of outcome indicators for the purpose of 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 student 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 intra-national level the debate has now shifted markedly from this simplistic assumption towards a recognition that institutional and subsystem comparisons must be contextualised, principally by making adjustments for student status and achievements on entry to the education system or particular phases of it (Sanders and Horn, 1994; FitzGibbon, 1992; DFE, 1995a).

In the UK the debate has centred around the notion of adjusted (‘value added’) comparisons among schools and other educational institutions. In a briefing paper, the Department for Education (DFE, 1995a) proposes that a ‘baseline’ of prior attainment should form the adjustment measure for any particular phase of schooling and points out that this will be a better reflection of a schools’ 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, while worthy and an improvement upon unadjusted ‘raw’ league tables, is generally unrealisable: 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 to 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.

- ***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 (NHS Management Executive, 1992) have been developed into a set of population outcome indicators for the NHS calculated at the purchasing authority level (McColl and Gulliford, 1993). These are now distributed in printed and disk form as the Public Health Common Dataset (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 (NHS 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 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 this procedure would also benefit from additional acknowledgement of uncertainty.

- ***Hospital trusts (the providers):***

Before the reorganisation of the NHS, most hospital outcome data was 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 (NHS Executive, 1995). There has been considerable public debate surrounding each such publication: institutions tend to be immediately ranked by the media and the apparent ”best” and ”worst” become subject to close scrutiny, accompanied by criticism from clinicians and statisticians of the naive 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" (BMJ, 1995). Past comparisons between the outcomes achieved by different hospitals in the UK have generally been strictly anonymised, 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 above, is already occuring in Scotland.

In the United States 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 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 extensive discussion of issues surrounding hospital comparisons, see a recent special issue of the Annals of Thoracic Surgery (1994).

- ***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 impact 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 adjusting. 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 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 impact 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 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 the latter 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, 1995) 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 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 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' ones 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 Generalised 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 exam score as response (Y). We can write a 2-level variance components model

$$\begin{aligned} y_{ij} &= \beta_0 + u_j + e_{ij} \\ \text{var}(u_j) &= \sigma_u^2 \\ \text{var}(e_{ij}) &= \sigma_e^2 \end{aligned} \qquad (1)$$

where y_{ij} is the exam score for the *i*-th student in the *j*-th school, u_j is the residual or 'effect' for the *j*-th school and e_{ij} the residual for the *i*-th student in the *j*-th 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 $\text{var}(\hat{u}_j)$ or alternatively $\text{rank}(\hat{u}_j)$ and var [rank($\hat{u}_j$)], which in turn can be used for comparisons among institutions. We shall discuss exactly how these can be used below. 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 3-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, for example 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 emphasise 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 summarised by, say, 95% intervals: maximum (quasi) likelihood models lend themselves to bootstrapping samples, while Markov chain Monte Carlo techniques (Besag et al., 1995; Spiegelhalter et al, 1995) easily accommodate the ranking of the set of parameter realisations 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 among 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 generalised 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 re-analysis 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 too are influenced by extra-institutional factors such as social background, prior behaviour and achievement, then they need to be contextualised 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 GCE examinations where earlier 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 (DFE, 1994, 1995b). We have chosen to use data from one local education authority but the points we wish to make will apply more generally.

For 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 recognise 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 recognised in the A Level Information System (ALIS, FitzGibbon, 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 individual department data 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 upon aggregating (1).

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

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 (1995) using student and school level data in a large scale analysis of A level results in 1993.

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

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

for the *i*-th student in the *j*-th (1,2) school in the *h*-th (1,2) year. This model is then aggregated to the school level as before. Thus u_{0j}, u_{1j} respectively are the *j*-th school (intercept) effect and difference between year 2 and year 1 with associated variances and covariance. The following two figures show the estimates, which are approximately independent, ordered for the set of schools: Figure 2 shows estimates of the school averages and overlap intervals and Figure 3 shows estimates of the year 2 - year 1 differences for each school with overlap intervals. The school numbers are displayed on each Figure.

(Figure 2 here)

(Figure 3 here)

In terms of pairwise comparisons, while 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. This latter result is also found to hold for trend estimates of up to five years (Gray et al, 1995).

Table 1 gives the parameter estimates from fitting model 3. We notice the imprecision of these estimates based upon a small number of schools.

Table 1. Parameter estimates from fitting model 3.

Fixed	Estimate	s.e.
Intercept	15.8	
Year	-1.4	0.7
Random between schools (variances and covariance, standard errors in brackets)		
	Intercept	Year
Intercept	4.06 (3.2)	
Year	-0.21 (2.5)	1.69 (4.3)

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 Figures 2 and 3 which compare institutions with one another are the appropriate ones 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 zero. 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 Figures 2 and 3 present a conservative picture being 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 using measures of intake. For example, in their classic paper ten years ago, Aitkin and Longford (1986) obtained a correlation of about 0.7 between 16 year examination results 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, 1995) 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 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) distinguish 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 among institutions. For such individuals they wish to ascertain the expected output achievement conditional 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 to 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 organisation 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 know *why* institutions differ. Nevertheless, the reasons for differences are of interest generally for accountability purposes. 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 feedback effects, namely that public knowledge of this kind which then is acted upon 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 (1995) based upon the analysis of A level and GCSE results for 325 schools and colleges and 21,654 students. Figure 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.

(Figure 4 here)

The correlation for Figure 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 comparison then failure to use this will, prima facie, result in a number of highly inaccurate inferences.

Figure 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.

(Figure 5 here)

From this figure 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 institutional differences will have a limited utility. A ranking such as that in figure 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 judgements upon individual institutions. The analysis of Gray et al (1995) which uses GCSE as outcome with adjustment for intake achievement at 11 years, confirms a similar picture for both a five year average and a five year trend.

4.3 Interpretation and impact

We have already demonstrated that, with current data, even after adjustment, finely graded comparisons among institutions are impossible. Nevertheless, it is possible that in certain circumstances we might be able to achieve better adjustments and hence more accurate comparisons. While this cannot be ruled out, and is certainly a legitimate area for further research, there do 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 upon 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 emphasising 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 a number of areas where it is fairly clear that current models are inaccurate. To begin with, measures of input used to date have been opportunistic in the sense that they happen to have been available. The use of verbal reasoning or reading achievement to adjust for overall GCSE 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, 1995) has shown that when GCSE is the output it is inadequate to adjust solely for 11 year achievement and that information about 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 predictor variables in these models. If this is done we would expect possibly substantial changes in parameter estimates and institutional comparisons (Woodhouse et al, 1996).

Any comparison between institutions on the basis of A level results is inevitably out of date. If adjustments are made using GCSE results then a comparison can apply, at the earliest, to a cohort about to enter an institution three 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 six years. If results are aggregated over, say, three years, then for A levels the lag is between three and five 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 well cause them to give only a small amount of weight to institutional performance comparisons 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 data appropriateness and quality 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

- ***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 getting to hospital in the first place, rather than the quality of care given once admitted.

- ***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 ICD9 codes, and the consistency of such ICD9 coding across institutions.

- ***The definition of the outcome:***

Thirty-day mortality is obtainable in the Scottish analysis due to their record linkage scheme, whereas in-hospital mortality is used in the US where such routine linked follow-up does not exist. The latter 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) .

- ***Selection and definition of confounder variables.***

Measures of severity of illness at admission to hospital have been criticised for not fully taking into account known discrepancies in outcomes associated with social background and other factors.

- ***Quality and completeness of data.***

McKee and Hunter (1995) identify problems with routine sources of adjustment data, while 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).

- ***Deliberate manipulation of data:***

This is covered in Section 5.3.

5.2 Statistical analysis and presentation

5.2.1. Models

The use of multi-level or hierarchical models has been pioneered in two areas closely related to institutional comparison. The first concerns the utilisation 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 U.S. More recently, Gatsonis et al (1993, 1995) employ random-coefficient logistic regression models in comparing rates of coronary angiography between U.S. 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 in order to obtain more accurate estimates of cancer incidence in small areas.

Examples of the use of multi-level models in institutional comparisons include Thomas et al (1994) in their analysis of mortality in Medicare patients, and Leyland et al (1995) when comparing length of stay in Scottish hospitals. Closest to our approach is Normand, Glickman and Gatsonis (1995), who include 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 zero to 5 'stars' to trusts with no comment on uncertainty (NHS Executive, 1995). Publications associated with the Public Health Common Dataset 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 disk does provide confidence intervals. Scottish data (Scottish Office, 1994) presents confidence intervals in all graphics, although smaller institutions are not shown in graphs. We note that one of the important consequences of using multi-level models should be that suitable adjustment is made for size and so large and small units can be simultaneously presented.

In order 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 using data from the Scottish outcomes study (Scottish Office, 1994) on teenage (ages 13-15) conception rates in the period 1990 to 1992 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 (NHS Management Executive, 1992). Figure 6a 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 multi-level 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, 100000) 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.

(Figure 6a here)

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 Figures 6b: 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 multi-level 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.

(Figure 6b here)

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 adjustment with regard to operative mortality. The multi-level model, in spite of making the individual estimates more accurate, has the effect of making the ranks even more uncertain.

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 exist (Iezzoni, 1994). Recent applications in which UK institutions have been explicitly (although anonymously) compared include survival of premature babies using the CRIB scoring system (de Courcy-Wheeler et al, 1995) and comparison of survival in intensive care units 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 surgery (CABG) seeks to create a cardiac profile system which assesses the performance of hospitals and surgeons over time, independent of the severity of individual patients' pre-operative 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 describe how the publication in December 1990 of a league table of hospital CABG-related mortality in the New York Times 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 1990-1992: 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 his patients. The ratio of observed to expected mortality can be interpreted as the surgeon’s standardised mortality rate, which when multiplied by the state-average mortality of 2.99% provides a ‘risk-adjusted mortality rate’ which forms the basis for comparisons between individuals.

Table 2. Observed, expected and risk adjusted surgeon mortality after coronary artery bypass graft surgery, 1990-1992 (Part of Table 4 of New York State Department of Health (1993), ranked by RAMR)

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

(Figure 7a here)

Figure 7a 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 of 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. Figure 7a also shows the effect on estimated mortality when assuming 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 having an interval that excludes that state average. Estimates are from a simulation with iterations and prior distributions matching those in the previous example.

(Figure 7b here)

Figure 7b shows the intervals for the rankings for the independent and multi-level model. It is clear that the intervals are very wide and for the independent estimates we can be confident about whether 5 surgeons lie in the upper or lower half: Green and Wintfeld (1995) use 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 performance. 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, availability of resources, throughput of patients and so on, 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) illustrate 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 randomisation is not considered feasible. McKee and Hunter (1994, 1995) provide a good discussion from a UK perspective, emphasising the limitations in data availability 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 quality of care. It is clear, however, that there are a number of techniques by which the results of such an exercise can be manipulated - this is known as ‘gaming’ in the US. An obvious example is provided by Green and Wintfeld (1995) in which the reported incidence of risk factors that would increase expected mortality rose after the introduction of the program: for example, reported congestive heart failure rose from 1.7% 1989 to 7.6% in 1991. McKee and Hunter (1995) point out that in the UK the imprecise description of a ‘finished consultant episode’ allows considerable scope for inflation of activity, while 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 depth the issue of data quality and appropriateness in both adjustment and outcome measures. Even where available measures are judged to be acceptable, however, there are inevitable limitations in making institutional comparisons 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 institutional comparisons, 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 among 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’ exam scores are used (DFE, 1995a).

This implies that current official support for output ‘league tables’, even adjusted ones, is misplaced and Government should be concerned that potential users are properly informed of their shortcomings. If such tables continue to be produced then they need an accompanying health warning about their use. Recently, the Department for Education and Employment (DFE, 1995a) has published analyses and charts of A level institutional performance indicators, adjusted for GCSE scores and using efficient multilevel modelling techniques. These analyses recognise differential effectiveness and are based upon 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 upon them an authority they do not possess.

An overinterpretation of a set of rankings where there are large uncertainty intervals, as in the examples we have given, can lead both to unfairness and 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 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 institutional comparisons 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 dissemination. There are certain kinds of institutional information which it may well be justified in disseminating widely. Information about the physical environment of a hospital or school or the quality of the organisation 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. On the other hand, there are some aspects of process 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 contextualised 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, 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, that is 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 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 (FitzGibbon, 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 with the remaining institutions being anonymised. While 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, for example by ignoring the existence of wide uncertainty intervals or by 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, while we have been generally critical of many current attempts to provide judgements about institutions, we do not wish to give the impression that we believe 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 criticise is that they distract both attention and resources from this worthier aim.

References

Aitkin, M. and Longford, N. (1986). Statistical modelling in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society*, A, **149**, 1-43

Annals of Thoracic Surgery (1994). Using Outcomes Data to Improve Clinical Practice: Building on Models from Cardiac Surgery. *Annals of Thoracic Surgery*, **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*. John Wiley and Sons: Chichester, England.

Besag, J., Green, P.J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion. Statistical Science, 10, to appear.

BMJ (1995) `Hospital league tables derided', *British Medical Journal*, **311**, 200.

Bumiller E (1995) `Death-rate rankings shake New York cardiac surgeons'. *New York Times*, September 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,** 671-81.

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. *Archives of Disease in Childhood*, **73** , F32-6.

Department of Health (1994) *Public Health Common Data Set 1993, England Volume 1*. Produced by the Institute of Public Health, University of Surrey.

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

DFE (1995a). *GCSE to GCE A/AS value added: briefing for schools and colleges*. London, Department For Education.

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

FitzGibbon C (1992*) School Effects at A level: Genesis of an Information System* in D Reynolds & P Cuttance (Eds) School Effectiveness Research Policy and Practice, 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. *Medical Care*, **33**, 625--642.

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

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398--409.

Gilks, W.R., Thomas, A., and Spiegelhalter, D.J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, **43**, 169--78.

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold, New York, Halsted Press.

Goldstein, H. and Healy, M.J.R. (1995). The graphical presentation of a set of means. *J. Royal Statistical Society, A.*, **158**, 175-77

Goldstein, H. and Sammons, P. (1995). The influence of secondary and junior schools on sixteen year examination performance. (Submitted for publication).

Goldstein, H. and Thomas, S. (1995). Using examination results as indicators of school and college performance. *J. Royal Statistical Society, A.* (to appear).

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 England Journal of Medicine*, **332** , 1229-32.

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. *Journal of American Medical Association*, **271** , 761-6.

Iezzoni L I (1994) (Editor) *Risk Adjustment for Measuring Health Care Outcomes*. Michigan: 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 Amer Med Assoc*, **260**, 3611--6.

Leyland, A.H. and Boddy, F.A. (1995) Measuring performance in hospital care: the example of length of stay in gynaecology. *European Journal of Public Health* (to appear).

Leyland, A.H., Pritchard, C.W., McLoone, P. and Boddy, F.A. (1991) Measures of performance in Scottish maternity hospitals. *British Medical Journal*, **303**, 389-393.

Lowry, S. (1988) Focus on performance indicators. *British Medical Journal*, **296**, 992--994.

McArdle, C.S. and Hole, D. (1991). Impact and variability among surgeons on postoperative morbidity and mortality and ultimate survival. *British Medical Journal*, **302**, 1501-1505.

McColl, A.J. and Gulliford, M.C. (1993) *Population Health Outcome Indicators for the NHS: a feasibility study*. Faculty of Public Health Medicine, 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.

McKee, M. and Hunter, D. (1995). Mortality league tables: do they inform or mislead? *Quality in Health 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 England Journal of Medicine*, **307**, 1310-1314.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications, (with discussion). *Journal of American Statistical Association*, **79**, 47-65.

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

NHS Executive (1995). *The NHS Performance Guide 1994-1995*. Leeds . NHS Executive .

NHS Management Executive (1992). *The Health of the Nation: A strategy for health in England*. London . HMSO.

NHS Management Executive (1993). *Local Target Setting - a discussion paper.*

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 USA.

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

Orchard, C. (1994). Comparing healthcare outcomes. *British Medical Journal,* **308**, 1493--6.

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. of Educational and Behavioural Statistics* (to appear).

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. *Journal of American Medical Association,* **264** , 1989--94.

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. *British Medical Journal*, **307**, 977--81.

Sanders, W. L. and Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): mixed model methodology in educational assessment. *J. of Personnel Evaluation in Education*, **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? *Journal American Statistical Association*, **90** , 7-18.

Smith, P. (1990). The use of performance indicators in the public sector. *Journal of the Royal Statistical Society,* 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? *British Medical Journal*, **311,** 151-55.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.30.* MRC Biostatistics Unit , Cambridge.

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. *Statistics in Medicine*, **13**, 889--903.

Woodhouse G. and Goldstein H. (l988). Educational Performance Indicators and LEA League Tables. *Oxford Review of Education* **14** 301-320

Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society*, A, 159 (to appear).

Yates, J.M. and Davidge, M.G. (1984) Can you measure performance? *British Medical Journal*, **288**, 1935--1936.

Legends for Figures

Figure 1: Simple regression relationships of ‘output’ on ‘input’ for two hypothetical institutions.

Figure 2: School intercept residual estimates and 95% overlap intervals.

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

Figure 4: Residual estimates for middle GCSE group by mean A level score (standardised scores).

Figure 5: Pairwise 95% overlap intervals based on adjusted residuals for middle GCSE group (standardised scores).

Figure 6a: Estimates and 95% intervals for teenage conception rates assuming independent (solid lines) and exchangeable (dashed lines) Scottish Health Boards: the English Health of the Nation target of 4.8 per 1000 is shown.

Figure 6b: Mean and 95% intervals for rank of Health Board assuming independent (solid lines) and exchangeable (dashed lines) Boards.

Figure 7a: Estimates and 95% intervals for risk-adjusted mortality rates assuming independent (solid lines) and exchangeable (dashed lines) surgeons: the state average of 2.99% is shown.

Figure 7b: Mean and 95% intervals for rank of surgeon assuming independent (solid lines) and exchangeable (dashed lines) surgeons.