

Papers

Was Rodney Ledward a statistical outlier? Retrospective analysis using routine hospital data to identify gynaecologists' performance

Mike Harley, Mohammed A Mohammed, Shakir Hussain, John Yates, Abdullah Almasri

Abstract

Objectives To investigate whether routinely collected data from hospital episode statistics could be used to identify the gynaecologist Rodney Ledward, who was suspended in 1966 and was the subject of the Ritchie inquiry into quality and practice within the NHS.

Design A mixed scanning approach was used to identify seven variables from hospital episode statistics that were likely to be associated with potentially poor performance. A blinded multivariate analysis was undertaken to determine the distance (known as the Mahalanobis distance) in the seven indicator multidimensional space that each consultant was from the average consultant in each year. The change in Mahalanobis distance over time was also investigated by using a mixed effects model.

Setting NHS hospital trusts in two English regions, in the five years from 1991-2 to 1995-6.

Population Gynaecology consultants (n = 143) and their hospital episode statistics data.

Main outcome measure Whether Ledward was a statistical outlier at the 95% level.

Results The proportion of consultants who were outliers in any one year (at the 95% significance level) ranged from 9% to 20%. Ledward appeared as an outlier in three of the five years. Our mixed effects (multi-year) model identified nine high outlier consultants, including Ledward.

Conclusion It was possible to identify Ledward as an outlier by using hospital episode statistics data. Although our method found other outlier consultants, we strongly caution that these outliers should not be overinterpreted as indicative of "poor" performance. Instead, a scientific search for a credible explanation should be undertaken, but this was outside the remit of our study. The set of indicators used means that cancer specialists, for example, are likely to have high values for several indicators, and the approach needs to be refined to deal with case mix variation. Even after allowing for that, the interpretation of outlier status is still as yet unclear. Further prospective evaluation of our method is warranted, but our overall approach may be potentially useful in other settings, especially where performance entails several indicator variables.

Introduction

The Ritchie report was based on one of the most detailed inquiries yet undertaken into the clinical practice of an individual gynaecologist, Rodney Ledward.¹ It focused on the clinical work of Ledward in the NHS and the private sector and examined allegations about failings in his practice. The criticisms made,

and subsequently substantiated, against Ledward included lack of care and judgment preoperatively, failings in surgical skills, inappropriate delegation to junior staff, and poor postoperative care and judgment.

In common with many other external and internal inquiries, little use was made of comparative data regarding the performance of individual consultants or surgical teams. For over 20 years, routine data sources such as the hospital episode statistics have been widely perceived as being of little value because of problems with completeness and accuracy, and it has been assumed that the type of information required to identify poor performance would necessitate a new data collection system. The Department of Health proposed the introduction of a "near miss" reporting system and dismissed the use of hospital episode statistics for identifying poor clinical quality, observing that historically, the uses of these data have concentrated on recording and assessing activity levels and on performance, including technical efficiency.² Much is of variable quality and equally variable relevance to the quality and outcomes of the care that the NHS provides.²

Despite these concerns, hospital episode statistics data were used in the Bristol inquiry,³ albeit not to study the work of individual surgeons or teams. The conclusion of the subsequent Kennedy report regarding hospital episode statistics was unequivocal; hospital episode statistics "was [sic] not recognised as a valuable tool for analysing the performance of hospitals. It is now, belatedly." This paper explores this theme, by comparing the performance of 142 gynaecology consultants with the performance of Ledward over a period of five years, to determine if Ledward was a statistical outlier according to hospital episode statistics data.

Methods

Disaster theory^{4,5} proposes that poor performance in an organisation usually manifests itself in several ways. Applying a mixed scanning approach,^{6,7} we sought to identify several measurable characteristics that might imply consistent failure in performance. Using the review of the Ritchie report, other reports of alleged malpractice, a general review of literature on performance failures, and discussions with a practising gynaecologist, we compiled a provisional list of 11 variables that could be indicative of poor performance and could be derived from hospital episode statistics. We refined this list by eliminating any variables that had high inter-correlations (for example, a multiple correla-



An appendix with statistical details is on bmj.com

Table 1 Seven clinically relevant indicator variables from hospital episode statistics

| Indicator | Reason for choice |
|--|--|
| % of finished consultant episodes with complications (recorded)* | High levels might be associated with poor surgical skills |
| Mean length of spell (days) | Long stay might be caused by high levels of complications |
| % of finished consultant episodes with more than two operations (recorded) | High proportion might be the result of poor surgical technique necessitating further surgery |
| % of finished consultant episodes where spell is longer than episode | High levels might imply complications requiring transfer to another specialist |
| % of finished consultant episodes for dilatation and curettage on women aged <40 | High proportion might imply inappropriate practice |
| % of finished consultant episodes for sterilisation on women aged <25 | High proportion might imply inappropriate practice |
| % of finished consultant episodes for hysterectomy on women aged <30 | High proportion might imply inappropriate practice |

*Complications were obtained by scanning all seven hospital episode statistics diagnostic fields for ICD-9 codes 996-999 and ICD-10 codes T80-T88: "Complications of surgical and medical care not elsewhere classified."

tion coefficient R^2 of over 0.2) with another variable. The selection of which indicator to retain was based on face validity. Furthermore we did not use mortality as one of our indicator variables because death in gynaecology is a rare event, and we were scanning for overall poor quality of care. We produced a list of seven indicator variables (table 1), largely on the grounds that they were clinically relevant (face validity) and seemed to have some directional properties, in that high values were in general likely to indicate poor performance. Nevertheless, we emphasise that, for each indicator, valid reasons may exist that could credibly explain performance occurring in the high end of that indicator distribution. However, what is considered much less likely is that the same team would display extreme performance across a basket of indicators. A team in this context refers to a single consultant and the junior doctors who deal with his or her patients.

We obtained complications by scanning all seven diagnostic fields of hospital episode statistics for *International Classification of Diseases*, 9th edition (ICD-9) codes 996-999 and ICD-10 codes T80-T88: "Complications of surgical and medical care not elsewhere classified."

We then calculated each indicator for each of the years from 1991-2 to 1995-6 for Ledward, his three colleagues in the same hospital, and all the gynaecologists in one other region, the West Midlands. The West Midlands data contained only anonymised consultant codes. At the time of our study, reliable data were not readily available for the whole of the region in which Ledward practised, so we were able to use the data only for Ledward's own hospital.

We undertook a retrospective desktop statistical analysis to determine whether Ledward could be identified as a statistical outlier. We assigned a study code to all consultants. Throughout the analysis, the analysts (SH and MAM) were blinded to the code of Ledward. The analysis proceeded in three stages.

Stage 1

Exploratory data analysis—In all, 143 consultants (coded 1-143) were in our data set, of whom 68 appeared in all five years. Table 2² shows the number of consultants in each year and the numbers excluded because of any missing data item. According to Little's D^2 statistic for missing data in multivariate data sets,⁸ the pattern of missing data was consistent with data missing at random ($P < 0.0005$).

Table 2 Numbers of consultants who were outliers at the 95% cut-off each year

| Year | Consultants in data set (consultants excluded*) | Consultants analysed | Consultant outliers (%) |
|--------|---|----------------------|-------------------------|
| 1991-2 | 98 (9) | 89 | 15 (17) |
| 1992-3 | 97 (5) | 92 | 15 (16) |
| 1993-4 | 104 (5) | 99 | 20 (20) |
| 1994-5 | 107 (6) | 101 | 16 (16) |
| 1995-6 | 117 (9) | 108 | 10 (9) |

*Excluded because of any missing data in that year.

Stage 2

We carried out a multivariate analysis to detect outliers, based on the computation of a robust Mahalanobis distance⁹ for each consultant in each year. The statistical details are provided in the appendix on bmj.com. For each year we computed, from the variable space of the seven indicators, a Mahalanobis distance for each consultant. The Mahalanobis distance is in essence a measure of the "distance" between the origin in the seven indicator variable space and a given data point. So a consultant with average values for each variable will have a Mahalanobis distance of zero, and this represents the origin. Consultants who are furthest away from the origin will have relatively larger distances. For each Mahalanobis distance we also derived an approximate 95% confidence interval, using computer simulation techniques. We randomly simulated each variable, for each consultant, 1000 times from an underlying binomial or normal distribution (the parameters of which were based on the observed data and the sample size). We used this simulated data set to derive 1000 simulated Mahalanobis distances for each consultant, which in turn were used to determine the approximate 95% confidence intervals for each consultant's distance.

The square root of the Mahalanobis distance ($\sqrt{\text{MD}}$) is known to follow approximately a \sqrt{k} distribution with k degrees of freedom (k being equal to the number of indicator variables, seven in our case),⁹ and so we used the mean of the \sqrt{k} , which is given by the \sqrt{k} degrees of freedom ($\sqrt{7} = 2.66$) to define outliers.⁹ Consultants with 95% intervals above the 2.66 threshold were deemed to be outliers. We report the number of outlier consultants for each year.

Stage 3

We also investigated the change in MD over the five years, using hierarchical analyses for repeated measurements. We constructed a two level hierarchical model, with consultant at level 1 (highest level) and their respective Mahalanobis distances at level 2 (lowest level). We used the standardised residual output from this model (see figure 2) to identify outliers beyond 2 standard deviations.

We used S-PLUS, version 6.1 (Insightful Corporation, Seattle, USA), with the Robust Library, version 1 (Beta II),¹⁰ and MLwiN, version 2.1c (University of London, London), for our analyses.

Results

Figure 1 shows the robust $\sqrt{\text{MD}}$ for each consultant for each year, and table 2 summarises the number of outlier consultants.

We also constructed a model to investigate the variation in $\sqrt{\text{MD}}$ over time (see bmj.com for further details), which reached significance ($P = 0.0043$). Figure 2 shows standardised residuals from the model. From this figure, we identified nine high outlier consultants and three low outlier consultants.

After these two analyses, MH revealed the consultant code and confirmed that Ledward was a statistical outlier (in three of

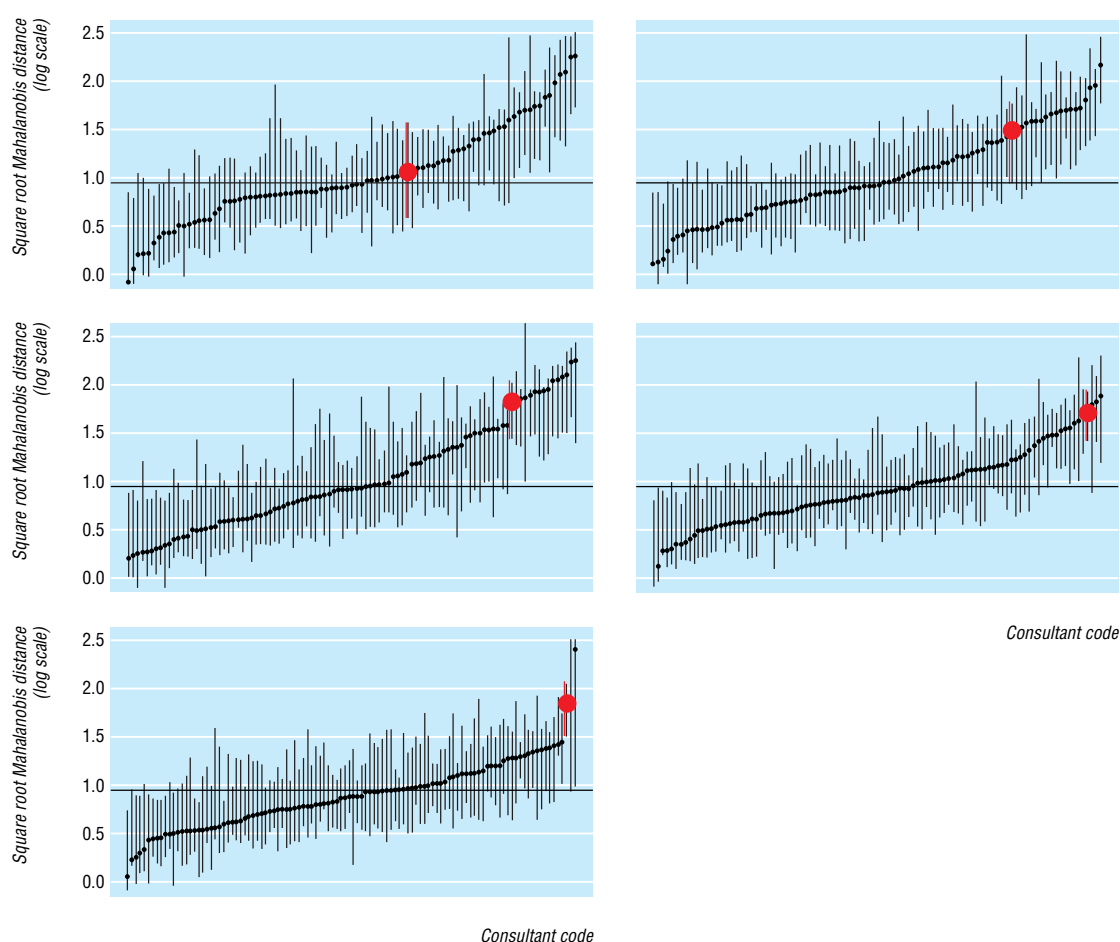


Fig 1 Plots showing the square root of the robust Mahalanobis distance (on the log_e scale to aid visualisation) for each consultant in each year (left to right: 1991-2, top left panel, and 1995-6, lower left panel). The horizontal line in each panel is the expected mean. Ledward is indicated by a filled circle. Vertical bars around each point are approximate, simulated, 95% intervals of uncertainty. Note that the ordering of the data in each panel is according to the y axis values, and so a given consultant will not necessarily appear on the same x axis value in each panel. This is illustrated by the filled circle for Ledward. To avoid confusion, we have therefore omitted the consultant codes from each plot

the five years of figure 1 and in figure 2). Figure 3 shows the variable values for Ledward. Several other consultants were outliers. Two consultants were outliers in all five years, two consultants were outliers in four years, and seven consultants (including

Ledward) were outliers in three years. Exploratory visual examination of the variable values for all these outlier consultants, also using figure 3 (results not shown) did not show any consultant as having consistently low values in all seven indicators.

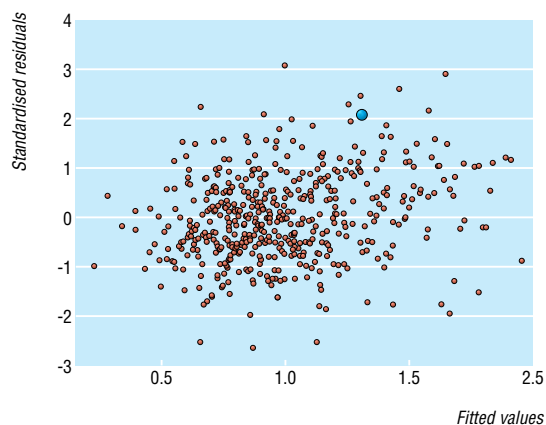


Fig 2 Fitted values versus the standardised residuals from statistical model. Consultants with standardised residuals outside the ± 2 standardised residuals envelope are deemed as outliers. Ledward is the larger filled circle

Discussion

Our study shows a robust statistical method for detecting outlier consultant firms, using a limited set of indicators derived from hospital episode statistics. In this post hoc analysis, Ledward was an outlier in three out of five consecutive years, and also when we considered the sequence of his Mahalanobis distances over time. Our method found other outlier consultants, but it was not part of our study remit to investigate them. Nevertheless, we strongly caution against over-interpreting the other outlier consultants as having “poor” performance because previous experience shows that we cannot reliably attribute residual unexplained variation to poor quality of care.¹¹ For example, our choice of indicators means that cancer specialists are likely to have values in the tails of our indicator distributions and so may appear as outliers in our analyses. The presence of an outlier should not invite capricious over-interpretation; it should, instead, invite us to undertake a rigorous scientific search for a credible explanation using a suitable model.¹¹

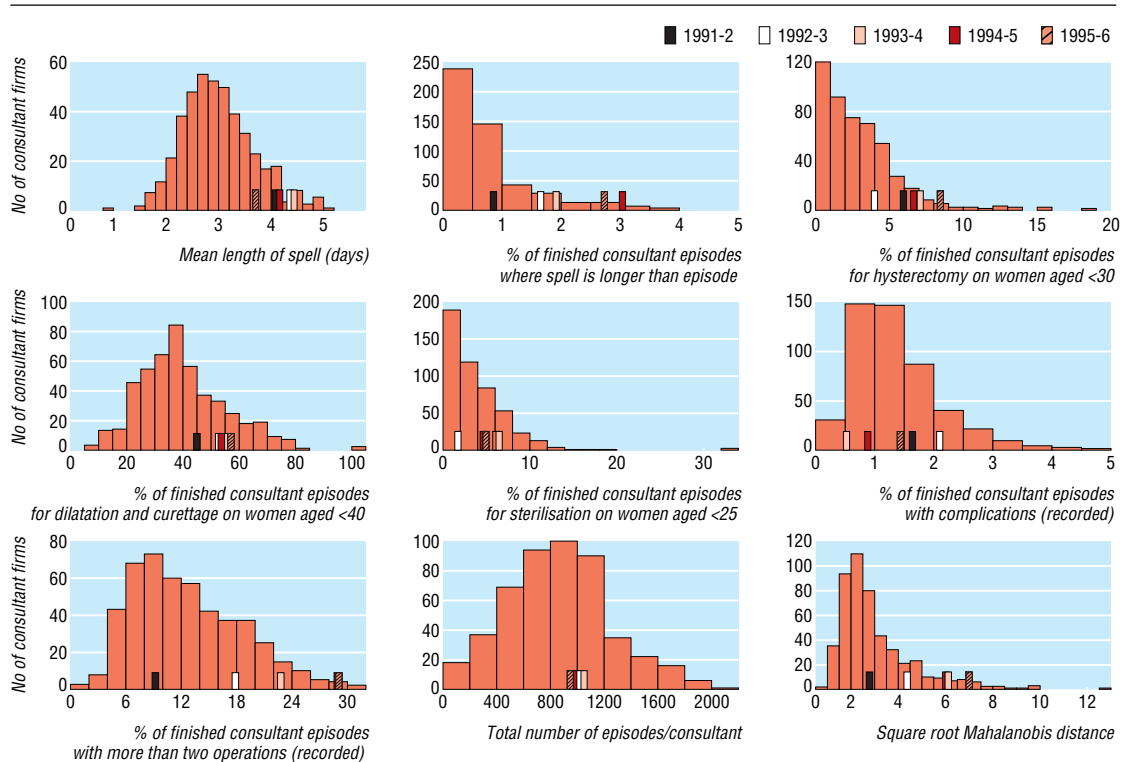


Fig 3 Histograms for the seven indicator variables, the total number of episodes per consultant, and the square root of the Mahalanobis distance for all years combined. Coloured boxes show the values for Ledward for each of the five years (1991-2 to 1995-6), respectively

Potential limitations of the study

The measurement of poor clinical performance in the NHS has no gold standard with which to compare this or any other statistical method,¹⁴ because in reality we are unable to calculate sensitivity and specificity of the “test” since we do not know the true underlying state of each subject. Recognising the limitations of statistics in this type of work is therefore important.¹⁴ Furthermore, the degree of statistical refinement applied to such problems must be weighed against the more fundamental limitations of the datasets available, their quality, and the role of human judgment in selecting the indicators.

Although we were not unduly hampered by the amount or pattern of missing data, the issue of what to do with subjects who have missing data is important. We excluded these subjects, but this creates the inappropriate impression that consultants with missing data may not be subject to a monitoring process. Although missing or poor quality data (an often cited criticism of hospital episode statistics data¹²) can hamper all analyses, they may not, as shown in the Bristol analysis,¹³ radically alter the ability to detect outliers. One statistical strategy to deal with missing data is imputation, although a more fundamental solution is to focus on the reasons for missing data or data of poor quality and deal with this through improved data collection methods as part of the overall monitoring system.¹⁴

The use of routine data sets such as hospital episode statistics places an important design constraint on analyses of this kind. Hospital episode statistics contain a limited number of variables, of which only some are potentially useful indicators of quality of care or of factors relating to the case mix of patients. However, this does not imply that analysis of routine data sets is without merit¹⁴; in recent years data from hospital episode statistics data have been used increasingly.^{3 15 16}

Furthermore, one can easily reduce or increase the number of statistical outlier signals by shortening or widening the inter-

vals of uncertainty, or by using non-robust statistical methods, but it is important to emphasise that this is not a purely statistical question. We must also consider the costs and benefits (including findings) of subsequent investigations. For example, after simulation to determine individual intervals of uncertainty, Ledward was an outlier in three of the five years, but his \sqrt{MD} was above the 95th centile (3.75) in four out of five years (fig 1), indicating that it may be prudent to review consultants with large Mahalanobis distance (say, above the 95th centile) even though the individual interval of uncertainty crosses (only just) the expected mean. So, although the setting of the threshold may be informed by statistical theory, we will ultimately require longer term empirical evidence to determine its utility.

Proposed framework for investigation

One proposed framework for investigation is the pyramid model of investigation.¹⁷ The model is based on the premise that the bulk of failure is attributable to the system and not the individual, and so the pyramid prescribes a check of the following variables in the order listed: check the data (recognising that some of the variation between consultants could simply be due to data quality or completeness^{13 14}), check the patient case mix, check the structure, check the process of care, and, finally, carefully check the carers involved. The pyramid model of investigation was applied recently in the case of two general practitioners who were identified via the Shipman inquiry as having “unacceptably” high death rates.¹⁷ These general practitioners were found to have large numbers of patients in nursing homes (a factor that was not taken into account in the underlying statistical model), and this credibly explained their high death rates.

Careful handling is essential

In responding to a signal of potentially poor performance we must be alert to some real dangers. For example, the presence of substantial criticism in the media, and even appearance before

What is already known on this topic

Routine hospital episode statistics have now been used to investigate mortality after cardiac surgery at hospital level (for example, in the Bristol inquiry) and more recently at consultant level

The use of hospital episode statistics data to identify broader suboptimal performance where death is a rare event remains less explored, especially at consultant level

What this study adds

A robust new method has been identified for scanning multi-indicator, multi-year data from hospital episode statistics to identify outlier consultants in Gynaecology

The method showed Rodney Ledward, who was the subject of the Ritchie inquiry, to be an outlier

the General Medical Council, does not guarantee that those so accused are actually guilty of poor performance,^{18 19} nor does it mean that all the remainder who have not been criticised are performing in an entirely acceptable manner. Once an individual has been publicly identified, the stigma remains,²⁰ and we cannot undo what has been done. These issues are especially important if the explanation for the poor performance is outside the gift of the individual carer.¹¹

Useful methods for monitoring performance

Although scanning methods¹⁴ such as ours will never have complete diagnostic certainty, they could be used to reliably identify signals from noise,¹³ which need to be systematically and sensitively examined, perhaps confidentially, by peers.²¹ Although our methods urgently need to be evaluated prospectively, organisations engaged in this type of performance monitoring, including the National Patient Safety Agency, the Healthcare Commission, the General Medical Council, the NHS Litigation Authority, and the National Clinical Assessment Authority may find our methods of interest. Nevertheless, although the ability to identify poorly performing clinicians after the event has its uses, prevention is preferable; but this presents an altogether different challenge—one that seeks to engineer the safety of patients into the process of care by design.

We thank J Duffy for his statistical advice at initial stages of this project; R Penketh, consultant gynaecologist, for his advice on indicators; and R Holder for his advice regarding the limits of uncertainty. We are grateful to S Evans and R Lilford for their critical comments on earlier drafts of the manuscript. Thanks are also due to the Kings Fund for funding the initial part of this work. AA is supported by the Swedish Foundation for International Cooperation in Research and Higher Education.

Contributions: The project team was headed by MH, who also carried out the preliminary analysis and wrote the first draft of the paper. JY secured

funding, undertook literature reviews, and was instrumental in the initial design. SH and MAM undertook the statistical analyses. MAM produced the final draft of the paper. AA, with guidance and support from SH and MAM, undertook the simulation work. All authors contributed to the writing of the final paper. MH is guarantor.

Funding: The Kings Fund funded the initial stages of this project.

Competing interests: None declared.

- 1 Department of Health. *The report of the inquiry into quality and practice within the National Health Service arising from the actions of Rodney Ledward. (The Ritchie report.)* London: Stationery Office, 2000.
- 2 Department of Health. *An organisation with a memory: report of an expert group on learning from adverse events in the NHS chaired by the chief medical officer.* London: Stationery Office, 2000.
- 3 Department of Health. *Report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995: Learning from Bristol. (The Kennedy report.)* London: Stationery Office, 2001.
- 4 Bignell V. *Catastrophic failures.* Oxford: Oxford University Press, 1977.
- 5 Turner BA. The organisational and interorganisational developments of disasters. *Admin Sci Q* 1976;21:378-97.
- 6 Etzioni A. Mixed-scanning: a "third" approach to decision-making. *Public Admin Rev*;27:385-92.
- 7 Yates JM. *The use of routinely collected information in the measurement of performance in the NHS.* Birmingham: University of Birmingham, 1986.
- 8 Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc* 1988;83:1198-202.
- 9 Rousseeuw PJ, Leroy AM. *Robust regression and outlier detection.* New York: Wiley, 1987.
- 10 Insightful Corporation. *S-PLUS 6 robust library user's guide version 1.0.* Seattle: Insightful Corporation, 2002.
- 11 Lilford RJ, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;363:1147-54.
- 12 Campbell SE, Campbell MK, Grimshaw JG, Walker AE. A systematic review of discharge coding. *J Public Health Med* 2002;23:205-11.
- 13 Aylin P, Alves B, Best N, Cook A, Elliot P, Evans SJ, et al. Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-96: was Bristol an outlier? *Lancet* 2001;358:181-7.
- 14 Spiegelhalter D, Murray G, McPherson K, Macfarlane A, Evans S, Curnow R, et al. *Monitoring clinical performance: a statistical perspective.* Submission to the Bristol Inquiry, 2002.
- 15 Jarman B, Gault S, Alves B, Hider A, Dolan S, Cook A, et al. Explaining differences in English hospital death rates using routinely collected data. *BMJ* 1999;318:1515-20.
- 16 Aylin P, Tanna S, Bottle A, Jarman B. Dr Foster's case notes: how often are adverse events reported in English hospital statistics? *BMJ* 2004;329:369.
- 17 Mohammed MA, Rathbone A, Myers P, Patel D, Onions H, Stevens A. An investigation into general practitioners associated with high patient mortality flagged up through the Shipman inquiry: retrospective analysis of routine data. *BMJ* 2004;328:1474-7.
- 18 Dunn PM. The Wisheart affair: paediatric cardiological services in Bristol, 1990-5. *BMJ* 1998;317:1144-5.
- 19 BBC News Online. Wisheart: callous or caring? <http://news.bbc.co.uk/1/hi/health/1124755.stm> (accessed July 2004)
- 20 BBC News Online. The second surgeon: Janardan Dhasmana. <http://news.bbc.co.uk/1/hi/health/1136419.stm> (accessed July 2004).
- 21 Mason S, Nicholl J, Lilford R. What to do about poor clinical performance in clinical trials. *BMJ* 2003;324:419-20.

(Accepted 20 January 2005)

doi 10.1136/bmj.38377.675440.8F

Inter-Authority Comparisons and Consultancy, Health Services Management Centre, University of Birmingham, Birmingham B15 2RT

Mike Harley *director*

John Yates *professor*

Department of Public Health and Epidemiology, University of Birmingham, Birmingham B15 2TT

Mohammed A Mohammed *senior research fellow*

Department of Primary Care and General Practice, University of Birmingham

Shakir Hussain *statistician*

Abdullah Almasri *visiting statistician*

Correspondence to: M Harley M.J.Harley@bham.ac.uk