

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 1996 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 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 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 hospital episode statistics have been widely perceived as being of little value because of problems with completeness and accuracy. 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 which concluded unequivocally: hospital episode statistics "was [sic] not recognised as a valuable tool for analysing the performance of hospitals. It is now, belatedly." This paper compares 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

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 those with high inter-correlations. We produced a list of seven indicator variables (table). Nevertheless, we emphasise that, for

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

BMJ 2005;330:929-32



Statistical details are on bmj.com



This is the abridged version of an article that was posted on bmj.com on 15 April 2005: <http://bmj.com/cgi/doi/10.1136/bmj.38377.675440.8F>

each indicator, valid reasons may exist that could explain performance occurring in the high end of that indicator distribution. Much less likely is that the same team would display extreme performance across a basket of indicators.

We obtained complications by scanning all seven diagnostic fields of hospital episode statistics. 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.

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, and the two analysts were blinded to the code of Ledward. The analysis proceeded in three stages.

Stage 1

Exploratory data analysis—Of the 143 consultants, 68 appeared in all five years. See bmj.com for the number of consultants in each year and the numbers excluded because of missing data. The pattern of missing data was consistent with data missing at random ($P < 0.0005$).

Stage 2

We carried out a multivariate analysis to detect outliers, based on the computation of a robust Mahalanobis distance⁴ for each consultant in each year. The statistical details are provided 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 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.

The square root of the Mahalanobis distance ($\sqrt{\text{MD}}$) is known to follow approximately a $\sqrt{\chi^2}$ distribution with k degrees of freedom (k being equal to the number of indicator variables, seven in our case),⁴ and so we used the mean of the $\sqrt{\chi^2}$, which is given by the \sqrt{k} degrees of freedom ($\sqrt{7} = 2.66$) to define outliers.⁴ Consultants with 95% intervals above the 2.66 threshold were deemed to be outliers. We report the number of outlier consultants for each year.

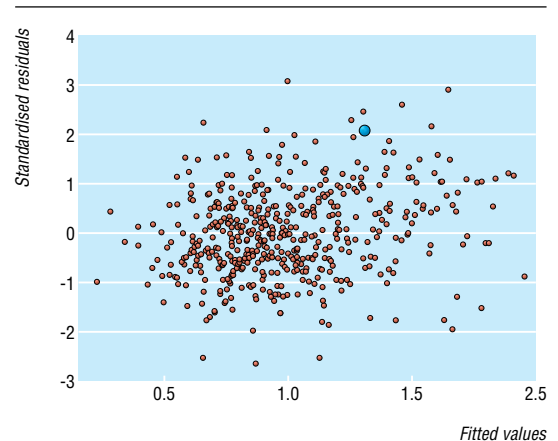


Fig 1 Fitted values versus the standardised residuals from statistical model. Consultants with standardised residuals outside the ± 2 standardised residuals envelope are deemed to be outliers. Ledward is the larger filled circle

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 1) to identify outliers beyond 2 standard deviations.

Results

See bmj.com for the robust $\sqrt{\text{MD}}$ for each consultant for each year, and summary of the number of outlier consultants. Ledward seemed to be an outlier in three out of five consecutive years

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 1 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. Figure 2 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.

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.”

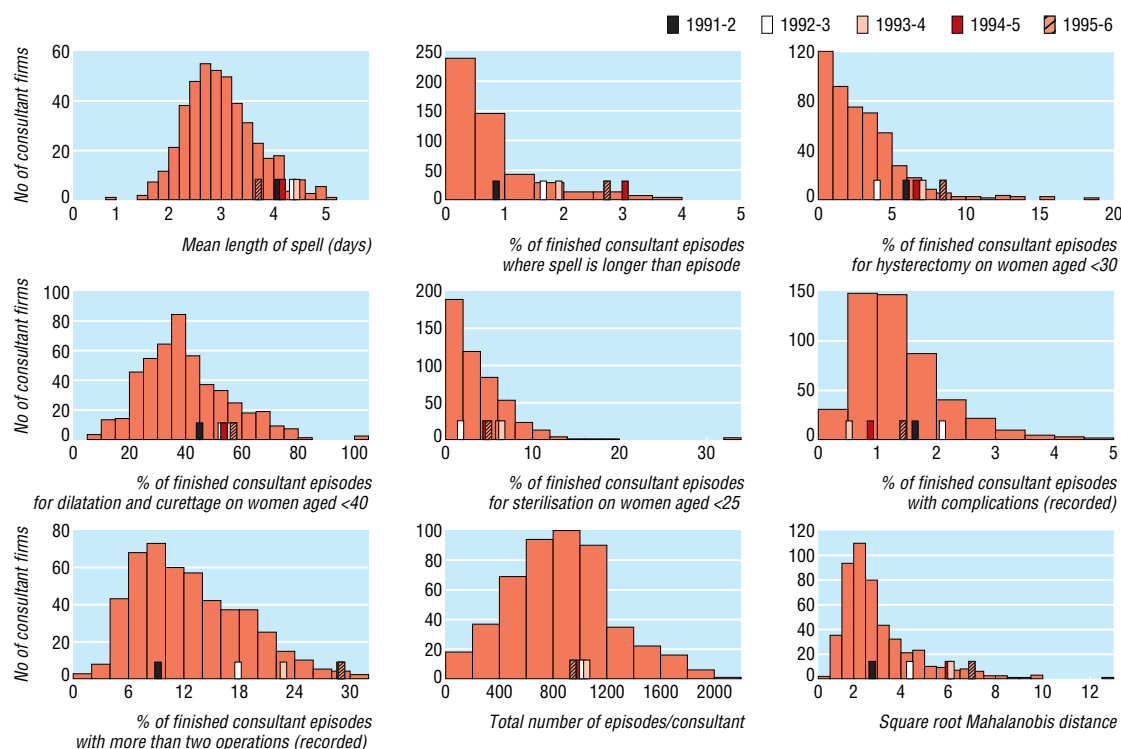


Fig 2 Histograms for the seven indicator variables, the total number of episodes per consultant, and the square root of the Mahalanobis distance for all years combined. Coloured boxes show the values for Ledward for each of the five years (1991-2 to 1995-6), respectively

Discussion

Our study shows a robust statistical method for detecting outlier consultant firms, using a limited set of indicators derived from hospital episode statistics. Ledward was an outlier in three out of five consecutive years, and also when we considered the sequence of his Mahalanobis distances over time. Other outliers should be regarded as signals meriting a scientific search for a credible explanation.⁵

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.⁶ Recognising the limitations of statistics in this type of work is therefore important.⁶ 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.

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 can hamper all analyses, they may not, as shown in the Bristol analysis,⁷ radically alter the ability to detect outliers. One statistical strategy to deal with missing data is imputation, although a more fundamental solution is to improve data collection methods.⁶

Hospital episode statistics contain a limited number of variables, of which only a portion are potentially useful indicators of quality of care or factors relating to the case mix of patients.

Furthermore, one can easily reduce or increase the number of statistical outlier signals by shortening or widening the intervals of uncertainty, or by using non-robust statistical methods, but 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 $\sqrt{\text{MD}}$ was above the 95th centile (3.75) in four out of five years, 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.

Proposed framework for investigation

The pyramid model of investigation⁸ is based on the premise that the bulk of failure is attributable to the system and not the individual. The pyramid prescribes a check of the following variables in the order listed: check the data, check the patient case mix, check the structure, check the process of care, and, finally, carefully check the carers involved.

Careful handling is essential

The presence of substantial criticism in the media, and even appearance before the General Medical Council, does not guarantee that those so accused are actually guilty of poor performance. Once an individual has been publicly identified, the stigma remains,⁹ 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.⁵

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)

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 was able to identify Rodney Ledward, who was the subject of the Ritchie inquiry

Useful methods for monitoring performance

Although scanning methods⁶ such as ours will never have complete diagnostic certainty, they could be used to reliably identify signals from noise,⁷ which need to be systematically and sensitively examined, perhaps confidentially, by peers. Prevention is preferable but this presents an altogether different challenge—engineering 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: See bmj.com

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 Rousseeuw PJ, Leroy AM. *Robust regression and outlier detection.* New York: Wiley, 1987.
- 5 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.
- 6 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.
- 7 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.
- 8 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.
- 9 BBC News Online. The second surgeon: Janardan Dhasmana. <http://news.bbc.co.uk/1/hi/health/1136419.stm> (accessed July 2004). (Accepted 20 January 2005)

doi 10.1136/bmj.38377.675440.8F

Incidence and risk factors for non-alcoholic steatohepatitis: prospective study of 5408 women enrolled in Italian tamoxifen chemoprevention trial

Savino Bruno, Patrick Maisonneuve, Paola Castellana, Nicole Rotmensz, Sonia Rossi, Marco Maggioni, Marcello Persico, Alberto Colombo, Franco Monasterolo, Donata Casadei-Giunchi, Franco Desiderio, Tommaso Stroffolini, Virgilio Sacchini, Andrea Decensi, Umberto Veronesi, for the Italian Tamoxifen Study Group

Abstract

Objective To assess the incidence, cofactors, and excess risk of development of non-alcoholic fatty liver disease, including non-alcoholic steatohepatitis, attributable to tamoxifen in women.

Design Prospective, randomised, double blind, placebo controlled trial.

Setting and participants 5408 healthy women who had had hysterectomies recruited into the Italian tamoxifen chemoprevention trial from 58 centres in Italy.

Intervention Women were randomly assigned to receive tamoxifen (20 mg daily) or placebo for five years.

Main outcome measure Development of non-alcoholic fatty liver disease in all women with normal baseline liver function who showed at least

two elevations of alanine aminotransferase (≥ 1.5 times upper limit of normal) over a six month period.

Results During follow up, 64 women met the predefined criteria: 12 tested positive for hepatitis C virus, and the remaining 52 were suspected of having developed non-alcoholic fatty liver disease (34 tamoxifen, 18 placebo)—hazard ratio = 2.0 (95% confidence interval 1.1 to 3.5; $P = 0.04$). In all 52 women ultrasonography confirmed the presence of fatty liver. Other factors associated with the development of non-alcoholic fatty liver disease

Correspondence to: S Bruno, Liver Unit, Azienda Ospedaliera Fatebenefratelli e Oftalmico, Corso di Porta Nuova 23, 20121 Milan, Italy. italysavino.bruno@fbf.milano.it

BMJ 2005;330:932-5

P+ Details of authors' affiliations are in the full version of the paper on bmj.com. Other members of the study group are listed on bmj.com.

ELPS This is the abridged version of an article that was posted on bmj.com on 21 March 2005: <http://bmj.com/cgi/doi/10.1136/bmj.38391.663287.E0>