Health Care Management Science 8, 237–241, 2005

Assessing Mortality Rates from Dubious Data—When to Stop Doing Statistics and Start Doing Mathematics

STEVE GALLIVAN*
Clinical Operational Research Unit, University College London, UK
E-mail: s.gallivan@ucl.ac.uk

Abstract. When trying to assess surgical outcomes at a particular centre, it is important to take account of case mix in terms of the types of operation performed. This is because those centres that undertake a disproportionately high number of complex operations might well be expected to have higher mortality rates than other centres whose case mix is more routine. From a statistical viewpoint, such case-mix adjustment is relatively straightforward if there are reliable risk estimates for different operation types. However this may not be the case and the risk estimates may have to be derived from several different sources which may not themselves be in agreement. Here, standard case-mix adjustment methods are no longer applicable and alternative analysis methods need to be used to make use of such unreliable risk estimates.

Keywords: estimation errors, mortality, optimisation

1. Introduction

This paper discusses one of the topics forming part of a keynote talk given at the fourth conference on quantitative modelling in the management of health care sponsored by the Institute of Mathematics and its applications. The talk was on the topic of "Measuring and improving performance in the NHS in the wake of Bristol". For readers unfamiliar with the nuances of UK health services, this title refers to a *cause célèbre* whereby a number of cardiac surgeons at the Bristol Royal Infirmary were suspended as the result of what were regarded as unacceptably high mortality rates in paediatric cardiac surgery. Subsequently, the government instituted a major public inquiry into the matter whose recommendations have since had a major impact on the operation of the NHS [1].

The author had a role to play in this affair. Prior to the unfolding of events at Bristol, he was part of a collaboration that developed an audit method known by the arcane and somewhat macabre acronym VLAD, (**V**ariable **L**ife **A**djusted **D**isplay) for assessing adult cardiac surgery outcomes taking into account variations in case-mix between different surgeons [2–4]. This is important to do since it is accepted that surgeons who deal with the riskiest cases can be expected to have higher than average mortality rates. The VLAD method was used as part of the investigation of Bristol carried out by the Royal College of Surgeons that was a precursor to the public inquiry. It has also become a standard audit method for adult cardiac surgery [5]. As a result of this work, the author was appointed to the expert group advising the Bristol Royal Infirmary Inquiry on 'Statistical' issues. This was a somewhat daunting experience since initially it was not clear what role there was on such an august committee for an operational researcher.

Much of the keynote talk concerned a review of VLAD and other mathematical methods for the analysis of surgical outcomes to detect anomalies that might indicate potentially suspect or declining performance, either on the part of a surgeon, the surgical team or within other parts of the health care process. This work has been amply discussed elsewhere [6] and it would be out of keeping to repeat this review here. Instead, this paper focuses on a particularly troublesome aspect of outcomes assessment that was encountered during the analysis associated with the Bristol case. This concerned anomalies in the national databases upon which the analysis relied. This was a particularly vexing issue since all concerned were well aware that applying textbook statistical methods relies on an assumption that the data being analysed are reasonably accurate and reliable. Unfortunately, this was patently not the case, and something else had to be dreamed up.

As is happened, the key notion was eventually expressible in fairly simple terms, which was a relief, since it was originally conceived in terms of a multi-dimensional optimisation problem, and there was thus the nightmare prospect of trying to find a way of conveying this to a lay audience.

Here, the analysis is presented in mathematical modelling terms, not only because this was the way that the problem was originally conceived, but also because this gives insight into potential extensions which may have a useful role in other contexts.

2. Discrepancies between national data bases

The main statistical analysis of outcomes was led by David Spiegelhalter [7], one of the UK's leading medical statisticians, and a central task of the expert panel was to provide a critical appraisal of this analysis.

An issue crucial to the analysis was that of case mix, since if it were true that surgeons at Bristol were performing operations that were more complex and demanding than at other UK centres, then it would not be surprising that they experienced higher mortality.

*Corresponding author.

Table 1
Comparison of data summaries for 1991–1994 related to congenital heart surgery taken from two national data sources (Taken from Spiegelhalter et al., 2001).

Operation category	Cardiac surgery register		Health episode statistics		Ratio of death rates
	Cases	Deaths	Cases	Deaths	
1	921	57	837	45	1.10
2	76	15	158	17	1.76
3	685	89	644	67	1.13
4	203	28	217	27	1.01
5	553	65	749	68	1.25
6	1525	11	1182	18	0.46
7	1141	26	1280	56	0.50
8	123	30	97	30	0.76
9	340	42	620	65	1.16
10	827	42	893	43	1.03
11	160	15	247	27	0.82
12	757	12	632	17	0.59

In order to examine the question of case mix, operations were divided into a number of different categories, each with broadly similar complexity. Analysis was carried out to establish typical mortality rates for these different categories of operation, in UK centres other than Bristol. Mortality data came from two separate sources: a major national database maintained by the UK Department of Health known as the health episode statistics (HES) database and a database maintained by UK cardiac surgeons known as the cardiac surgical register (CSR).

A major methodological problem was encountered in that it was found that there were major differences between the HES and CSR databases both in terms of the reported frequency of different types of operation performed and their outcomes. This is summarised in table 1.

Given this degree of divergence between the two data sources, no amount of sophisticated statistical analysis could provide convincingly accurate estimates for the mortality rates for the different categories of operation, yet such mortality estimates were central in order to take account of case-mix at Bristol.

3. Changing the focus of the analysis

To consider this analysis problem in mathematical terms we first introduce some notation and, since there is nothing to lose and possibly something to gain, we do this in rather general terms.

Suppose there are $(H + 1)$ surgical centres, indexed $0, \ldots, H$ and that we are interested in comparing mortality outcomes at centre 0 with the other H centres which will be termed *comparator centres*.

Let us suppose that there are C different categories of operation and D different data bases.

For $1 \leq c \leq C$, $1 \leq d \leq D$ and $0 \leq h \leq H$, let $N_{c,d,h}$ denote the number of cases of operation category c reported to have been carried out at centre h according to data base d and let $M_{c,d,h}$ denote the corresponding number of deaths recorded.

From this one can generate estimates of the mortality rates for different operation categories. For $1 \leq c \leq C$ and $1 \leq d \leq D$, let $\alpha_{c,d}$ be defined by

$$\alpha_{c,d} = \frac{\sum_{h=1}^{H} M_{c,d,h}}{\sum_{h=1}^{H} N_{c,d,h}} \tag{1}$$

which represents the mortality rates for the c-th operation category estimated from the d-th data base for each of the comparator centres.

There are various ways in which the outcomes at centre 0 could be judged, perhaps the most natural being to consider the ratio of the number of deaths at centre 0 with an estimate of the number of deaths that would be expected given the case mix; however it is mathematically more convenient to deal with the reciprocal that we shall refer to as the at the performance quotient

$$Q = \frac{\textit{Expected number of deaths at centre } 0}{\textit{Actual number of deaths at centre } 0} \tag{2}$$

a value of 1 being 'par for the course', lower values being more concerning.

If there are serious discrepancies between estimates from the different data bases, as is the case in table 1, this casts considerable doubt on the credibility of the data sources and no amount of sophisticated statistical analysis can confirm which, if any, is the data source that is most believable. Given this, there are inevitably problems generating a credible estimate for Q from the data available and attempting to do so is perhaps something of a futile exercise. Given this, a change of focus is sensible. Rather than regarding the assessment of centre 0 as a statistical problem concerned with estimating the quotient Q as 'accurately' as possible, one can accept that this may not be possible and instead ask a different question.

Central change of focus

Instead of

"*Can one estimate Q accurately*?"

change question to

"*Can one make a quantitative statement about Q that is credible*?"

Transforming the problem so that it is no longer so overtly statistical in nature considerably simplifies life and, in the context of a legal inquiry probably comes closer to providing the sort of evidence that is required. After all, the precise degree of excess mortality is somewhat besides the point, more important is to establish that it is very credible that there was indeed excessive mortality to a degree that was substantial.

4. Credible lower bounds for the performance quotient

With this new focus one can derive various mathematical expressions that might be thought of as credible bounds for the performance index Q, so for example for $1 \le d \le D$, we can define

$$\hat{Q}_d = \frac{\sum_{c=1}^{C} \alpha_{c,d} N_{c,d,0}}{\sum_{c=1}^{C} M_{c,d,0}} \tag{3}$$

being an estimate of the performance quotient according to the d-th data base. Hence one quantitative expression of interest is given by

$$\hat{Q}_{\max} = \max\{\hat{Q}_d | 1 \le d \le D\} \tag{4}$$

which is in some sense an upper bound for Q, the true performance quotient. Thus if we are prepared to believe that at least one of the D databases is fairly reliable, then it is credible that the true value of Q must be at most of the same order as $\hat{Q}_{\max}$.

Of course a sceptic, or indeed a defence lawyer, might complain that there is no evidence that any of the data bases are reliable, thus why should one base analysis on the assumption that at least one of them is.

In view of this, an alternative method for deriving a credible lower bound for the true value of Q can be framed in terms of the solution to a different optimisation problem. Let us suppose that for $1 \le d \le D$, we define functions as follows:

$$q_d(x_1, \ldots, x_C) = \frac{\sum_{c=1}^{C} x_c N_{c,d,0}}{\sum_{c=1}^{C} M_{c,d,0}} \tag{5}$$

being an estimate of the performance quotient according to the case mix based on the d-th data base, but with variables $\{x_1, \ldots, x_C\}$ representing the expected death rates for each operation class.

It is also a useful mathematical device to define a function in terms of the point-wise maximum values of the D different functions, thus we have

$$\breve{Q}(x_1, \ldots, x_C) = \max\{q_d(x_1, \ldots, x_C) | 1 \le d \le D\} \tag{6}$$

We can now construct $\tilde{Q}$, another conservative upper bound for the true performance quotient Q, in terms of the following optimisation problem:

$$\breve{Q} = \max Q(x_1, \ldots, x_C)$$

subject to

$$\min\{\alpha_{c,d} | 1 \le d \le D\} \le x_c \le \max\{\alpha_{c,d} | 1 \le d \le D\}, \quad 1 \le c \le C. \tag{7}$$

Setting aside the technical details of how to solve this optimisation problem for a moment, let us consider its construction. Let us suppose that we accept that we do not know any of the mortality rates for the different operation types (the x-values). For each, D estimates are available, one from each data base. Suppose that we are willing to accept that the true mortality rate for a given operation will be bracketed by the D estimates available. In that case, the value $\tilde{Q}$ gives an upper bound for what we would accept the true value of the performance quotient for the centre under scrutiny. Note that we are not assuming that any one data base is the most accurate nor that any one data base always gives the most appropriate mortality estimate.

Now it may be that we wish to extend this notion since we may want to take into account knowledge of what gives rise to discrepancies between data bases. For example, it may be that the overall number of cases and numbers of deaths in each data base are comparable, but that there are differences in how they have been apportioned amongst the different operation categories. A plausible reason for this would be the difficulty faced by medically unqualified coding clerks determining how to classify a particular case. In view of this, the optimisation problem (7) might be modified to give another credible upper bound Q' for the performance quotient at the centre under scrutiny:

$$Q' = \max \breve{Q}(x_1, \ldots, x_C)$$

subject to

$$\min\{\alpha_{c,d} | 1 \le d \le D\} \le x_c \le \max\{\alpha_{c,d} | 1 \le d \le D\}, \quad 1 \le c \le C$$

$$\min\left\{ \sum_{c=1}^{C} \sum_{h=1}^{H} M_{c,d,h} \,\middle|\, 1 \le d \le D \right\} \le \sum_{c=1}^{C} \sum_{h=1}^{H} x_c N_{c,d,h}, \quad 1 \le d \le D,$$

$$\sum_{c=1}^{C} \sum_{h=1}^{H} x_c N_{c,d,h} \le \max\left\{ \sum_{c=1}^{C} \sum_{h=1}^{H} M_{c,d,h} \,\middle|\, 1 \le d \le D \right\} \quad 1 \le d \le D. \tag{8}$$

Here, the final two sets of constraints ensure that whatever values are assigned to the mortality rates $\{x_1, \ldots, x_C\}$ then the total deaths estimated for the comparator centres are of the right order.

If one knew even more about the nature of errors in the databases, for example the probability that a particular form of miscoding would occur, then even more elaborate objective functions and constraints can be framed. Details of this are omitted since hopefully the principle is clear.

5. The credibility of case load estimates

Of course a tacit assumption in all these methods for generating credible upper bounds is that there is a belief that at least one of the data bases gives a credible estimate for the actual case load at the centre under scrutiny. This is a central issue since such estimates determine the denominator of the performance quotient and thus if all the estimates of case load are grossly in error, then this would be inherited by any of the bounding estimates for Q given by (4), (7) or (8).

In such circumstances it may be deemed necessary to derive credible estimates of the case load at the centre under scrutiny independently from the estimates in the data bases. Let us suppose this is the case and that for $1 \leq c \leq C$, that R_c denotes the number of operations of category c that are believed to have taken place at the centre under scrutiny during the period under review.

Let us suppose that we define a function as follows:

$$q'(x_1, \ldots, x_C) = \frac{\sum_{c=1}^{C} x_c R_c}{\sum_{c=1}^{C} R_c} \tag{9}$$

where the $x_1, \ldots, x_C$ represents unknown mortality rates for each of the different operation categories. This has a similarity to the surrogate performance quotient functions defined in (5), although here the case load at the centre under scrutiny is assumed to be known.

Mimicking the analysis carried out to derive the optimisation problem (7), we can derive Q'', another upper bound for the performance quotient Q as the solution to the following optimisation problem.

$$Q'' \quad \max q'(x_1 \ldots, x_C)$$

subject to

$$\min\{\alpha_{c,d} \mid 1 \leq d \leq D\} \leq x_c \leq \max\{\alpha_{c,d} \mid 1 \leq d \leq D\}, \quad 1 \leq c \leq C. \tag{10}$$

Again, this assumes that we are prepared to believe that the true mortality rate for a given operation will be bracketed by the D estimates available. Also, again, as with (7), we are not assuming that any one data base is the most accurate nor that any one data base always gives the most accurate mortality estimates

Further, as with (8), if the errors in the data bases are believed mostly due to miscoding of operation types, then we have a further upper bound for the performance quotient.

$$Q''' = \max q'(x_1 \ldots, x_C)$$

subject to

$$\min\{\alpha_{c,d} \mid 1 \leq d \leq D\} \leq x_c \leq \max\{\alpha_{c,d} \mid 1 \leq d \leq D\}, \quad 1 \leq c \leq C.$$

$$\min\left\{ \sum_{c=1}^{C} \sum_{h=1}^{H} M_{c,d,h} \,\middle|\, 1 \leq d \leq D \right\} \leq \sum_{c=1}^{C} \sum_{h=1}^{H} x_c N_{c,d,h}, \quad 1 \leq d \leq D,$$

$$\sum_{c=1}^{C} \sum_{h=1}^{H} x_c N_{c,d,h} \leq \max\left\{ \sum_{c=1}^{C} \sum_{h=1}^{H} M_{c,d,h} \,\middle|\, 1 \leq d \leq D \right\}, \quad 1 \leq d \leq D. \tag{11}$$

6. Solving the optimisation problems derived

The discussion so far has shown that there are several methods (4), (7), (8), (10) and (11), that can be used to express credible upper bounds for the performance quotient of a centre under scrutiny depending what assumptions are viewed as acceptable. Each depends on an optimisation process. As yet, little has been said about the technical details of how these optimisation problems should be solved, deliberately so, since the emphasis has been to motivate a different and non-statistical way of thinking about how to estimate credible upper bounds for the performance quotient.

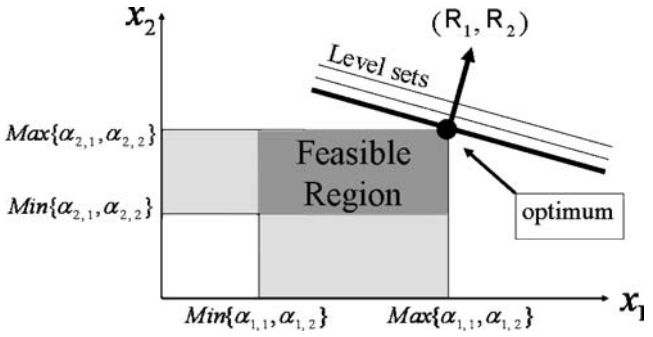

Figure 1. An illustration of the linear programming problem that can be used to derive the optimal choice of mortality rates in the case where there are two operation categories and two data bases.

In fact the technical issues associated with deriving optimum solutions simplify considerably in most cases.

The bounding estimate given by (4) is trivial to compute, since one merely needs to calculate a list of estimates and choose the largest. The other methods are apparently more complex since they require the solution of a constrained optimisation problem, however most simplify considerably.

This is easiest to discuss in terms of the optimisation problem (10) that is in fact a linear programming problem given the nature of the objective function $q'(x_1, \ldots, x_C)$, given by (9). Furthermore, the linear programming problem has a particularly simple geometric form, as illustrated in figure 1, in the case where one is dealing with just two categories of operation and two databases upon which to base mortality estimates.

In general, the feasible region is a convex polytope whose faces are parallel to those of the unit cube in C-dimensional Euclidean space. The gradient vector of the function $q'(x_1, \ldots, x_c)$ is simply $\nabla q'(x_1, \ldots, x_C) = (R_1, \ldots, R_C)(\sum_{c=1}^{C} R_c)^{-1}$ and this is normal to the level sets of $q'(x_1, \ldots, x_C)$, pointing in the direction of increase of the function. Now since all the components of the gradient vector are non-negative, the maximum value of $q'(x_1, \ldots, x_C)$ is attained at the vertex of the feasible region the sum of whose coordinates are maximum, since at this point, the corresponding level set will be tangential to the feasible region.

A similar argument can be used to simplify the optimisation problem (7). Each individual function $q_d(x_1, \ldots, x_C)$ is a linear function of $x_1, \ldots, x_C$ with non-negative coefficients. Thus, as with the analysis of problem (10), each of these linear functions will attain their maximum values at the vertex of the feasible region the sum of whose coordinates are maximum. Thus the maximum value $\tilde{Q}$ can be obtained by simply

evaluating the D functions $q_d(x_1, \ldots, x_C)$ at that point and choosing the maximum.

For the optimisation problem (11), the presence of the latter two sets of constraints complicates the geometric structure of the feasible region although it is noted in passing that given that the coefficients of the objective function are non-negative, the penultimate set of constraints are redundant, leaving a fairly simple linear programming problem.

Turning to the solution to optimisation problem (8), the solution can be found from the following equivalent optimisation problem whose solution simply requires the solution of a set of D linear programming problems:

$$Q' = \max\{Q''_d(x_1, \ldots, x_C) \mid 1 \le d \le D\}$$

where, for $1 \le d \le D$,

$$Q''_d = \max\{q_d(x_1, \ldots, x_C)\}$$

subject to

$$\min\{\alpha_{c,d} \mid 1 \le d \le D\} \le x_c \le \max\{\alpha_{c,d} \mid 1 \le d \le D\}, \quad 1 \le c \le C$$

$$\sum_{c=1}^{C}\sum_{h=1}^{H} x_c N_{c,d,h} \le \max\left\{\sum_{c=1}^{C}\sum_{h=1}^{H} M_{c,d,h} \,\middle|\, 1 \le d \le D\right\}, \quad 1 \le d \le D. \tag{12}$$

7. Use of such methods in relation to the bristol inquiry

Fortunately for the analysts, in the case of the Bristol Inquiry, the nature of the data available, and the scale of the mortality figures being analysed meant that it was unnecessary to go so far as to make use of the estimation procedures given by (11) or (12). This was something of a relief, since this avoided the necessity of implementing an algorithm to derive these optimum values. Equally, it avoided the problem of finding a way to explain these higher dimensional geometric arguments in the layman's terms required for evidence supplied to the Bristol Royal Infirmary Inquiry.

In the event, it was sufficient to explain that for each of the categories of operation, two mortality estimates were available, one based on the HES data base and the other based on CSR data base. From these, the combination of mortality rates, chosen to show Bristol in the best possible light, still gave rise to mortality estimates that were alarmingly high. This extreme 'sensitivity analysis' was incorporated in Spiegelhalter's statistical evidence and helped to establish beyond reasonable doubt that the statistical conclusions were sound. It was only in retrospect that the author can smile at the amount of time and mathematical energy expended to achieve such a simple analysis recommendation.

References

[1] The Inquiry into the management of care of children receiving complex heart surgery at the Bristol Royal Infirmary, http://www.bristol-inquiry.org.uk (2000).

[2] J. Lovegrove, O. Valencia, T. Treasure, C Sherlaw-Johnson and S. Gallivan, Monitoring the result of cardiac surgery by variable life adjusted display (VLAD), Lancet 350 (1997) 1128–1130.

[3] J. Lovegrove, C. Sherlaw-Johnson and S. Gallivan, Monitoring the performance of cardiac surgeons, Journal of the Operational Research Society 50(7) (1998) 684–689.

[4] C. Sherlaw-Johnson, J. Lovegrove, T. Treasure and S. Gallivan, Likely variations in perioperative mortality associated with cardiac surgery: When does high mortality reflect bad practice? Heart 84 (2000) 79–82.

[5] B. Keogh and R. Kinsman, National adult cardiac data base report 1990–2000 The Society of Cardiothoracic Surgeons of Great Britain and Ireland, 2000.

[6] T. Treasure, O.Valencia , C. Sherlaw-Johnson and S. Gallivan, Surgical performance measurement, Health Care Management Science 5 (2002) 243–248.

[7] D.J. Spiegelhalter, S. Evans, P. Aylin, and J. Murray, Overview of statistical evidence presented to the Bristol Royal Infirmary Inquiry concerning the nature and outcomes of paediatric cardiac surgical services at Bristol relative to other specialist centres from 1984 to 1995, *in: The Bristol Royal Infirmary Inquiry. Learning from Bristol. The Report of the Public Inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995* (The Stationery Office, London, Annex B, 2001).