EDUCATION CORNER

Standardized mortality ratios

Paul Taylor

Centre for Health Informatics and Multiprofessional Education, University College London, Stephenson Way, London, NW1 2HE, UK. E-mail: p.taylor@ucl.ac.uk

Accepted 16 September 2013

Mortality rates are increasingly used to compare the performance of hospitals and of units within hospitals. Different approaches have been taken to develop ‘standardized’ measures which adjust for the ‘expected’ mortality based on key characteristics of the population being treated. This article discusses the original motivation for looking at hospitals in this way in the UK, describes two recent incidents in which such measures were a critical element in decision making and reviews some of the strengths and weaknesses of the approaches. The current practice on the use of standardized mortality ratios is described.

Introduction

In recent months there have been numerous news stories in the UK about failing, or apparently failing, hospitals: some following on from the publication of the Francis Report into the appalling care received by patients at the Mid Staffordshire NHS Trust, others from the decision to suspend paediatric cardiac surgery at Leeds General Infirmary. The situations in the two trusts are very different, but in each the quality of care by was assessed using standardized measures of mortality. Different forms of such measures are increasingly used, in health systems around the world, as measures of quality.

The Bristol Royal Infirmary scandal: using national datasets to compare mortality rates

The use of standardized mortality rates to compare the performance of clinical teams first came into prominence in the UK in the wake of a scandal surrounding the deaths of children at Bristol Royal Infirmary between 1991 and 1995. That scandal emerged when a consultant anaesthetist collected data that he believed showed the mortality rates of his surgical colleagues to be excessive, but he was unable to persuade the hospital to take action. The facts of the story were leaked and led to a public inquiry chaired by Ian Kennedy QC and supported by a team that included Brian Jarman, a professor of primary care. The Inquiry commissioned statistical analysis from experts, including David Spiegelhalter, now Winton Professor for the Public Understanding of Risk at Cambridge, and Paul Aylin, now the Deputy Director of the Dr Foster Unit at Imperial College.

It might be thought that the statisticians’ task was straightforward, given that the statistic that matters here, the mortality rate, is simply the number of patients who died after treatment divided by the number of patients who were treated. The problem is that the Bristol mortality rate is of no interest on its own; what matters is how it compares with that of other centres. This means, at the very least, that the definitions of the two numbers have to be applied consistently across all the centres.

There were two national datasets available to the Bristol statisticians, both problematic. Hospitals return to the NHS a measure of activity known as hospital episode statistics or HES data. After a patient is discharged from hospital, the clinicians’ notes for that patient are read, often with some difficulty, by a ‘clinical coder’ who ‘codes’ the episode, matching the doctors’ narrative against two predefined lists of standard terms. The International Classification of Diseases (ICD) is used to assign a code for the primary diagnosis and for any co-morbidities. The Office of Population Censuses and Survey’s Classification of Interventions and Procedures (OPCS) is used to assign a code for the treatment or treatments received. The use of standard codes for both diagnoses...
and treatments make HES data appealing to statisticians. The process of coding from discharge summaries is, however, susceptible to error. In addition, although in-hospital deaths are recorded in HES, deaths that occur shortly after discharge are not.

The other national dataset available at the time was the Cardiac Surgical Register (CSR) compiled by the Society of Cardiothoracic Surgeons of Great Britain and Ireland. It included anonymized data on activity and mortality rates over the period dealt with by the Bristol Inquiry. Although such clinical audit databases are in many ways a more direct measure of clinical outcomes than HES data, comparisons have shown under-recording of cases. The Bristol team found that centres varied in: what staff they included; the sources of data used for reporting; and the definitions applied. The classification of complex diagnoses was a particular issue and it was sometimes hard to identify the surgical procedure performed.

The statisticians working for the Bristol Inquiry carried out analyses using both datasets; one could compare the mortality rate observed at Bristol directly with that of the other centres. However, the surgeons at Bristol admitted that their mortality rates were high but argued that this was because they undertook more difficult cases. The statisticians therefore built a logistic regression model to determine the expected mortality given the mix of cases seen at a centre, and calculated the excess deaths (actual minus expected deaths). This analysis is easy enough to perform repeatedly and so was used to assess the sensitivity of the statisticians’ calculations by carrying them out multiple times, varying those of the input values that might be susceptible to error. This regression model, however, although it takes account of the variation in the risk between different procedures, imposes a strong assumption: that the risk of a procedure is the same in different centres. A more sophisticated but time-consuming analysis was performed which avoided this assumption although, inevitably, it required assumptions about the variance that might be found between centres. From these analyses, both the CSR and the HES datasets showed evidence of excess mortality from 1991 to March 1995 in open operations in children under 1 year old. In this period, for these children, the mortality rate in Bristol was around double that of other centres. The estimated excess mortality was 19 out of the 43 deaths reported to the CSR and 24 out of the 41 deaths recorded in the HES. The results were taken as conclusive proof that mortality rates at Bristol were ‘divergent’.

Dr Foster and hospital standardized mortality ratios

Jarman had, by the time the Bristol Inquiry finished, already begun to look at using HES data to measure the performance of hospitals. In 2000 he and Sunday Times journalist Tim Kelsey persuaded the government to allow them to publish performance tables of NHS trusts. Jarman and Aylin founded the Dr Foster Unit at Imperial College to analyse HES data. Kelsey and others founded Dr Foster Intelligence in 2001 to publish the Good Hospital Guide and provide related commercial services to NHS Trusts. At the heart of both initiatives was the hospital standardized mortality ratio (HSMR), devised by Jarman and Aylin:

$$\text{HSMR} = \frac{\text{Actual deaths} - \text{expected deaths}}{\text{expected deaths}} \times 100.$$ 

Data on actual deaths were taken from HES data and restricted to in-hospital deaths. A logistic regression model was used to calculate the risk of death for the 50 most common diagnoses, which account for over 80% of admissions, based on a set of factors: sex, age, admission method (non-elective or elective), socio-economic deprivation quintile of the area of residence of the patient, diagnosis/procedure subgroup, co-morbidities, number of previous emergency admissions, year of discharge, month of admission and source of admission, and the use of the ICD code for palliative care. Using data on the mix of these factors seen by each trust, they calculated the expected death rate. If the expected death rate equalled the actual death rate, the trust had a score of 100. Scores tended to vary between 75 and 120.

Hospital standardized mortality rates at Mid Staffordshire NHS Trust

Between 1997 and 2008, the HSMR of Mid Staffordshire NHS Trust never got below 108. In all but 2 years, the 95% confidence interval around the HSMR was entirely above 100.

In 2007, the DFI’s Good Hospital Guide gave Mid Staffs an HSMR of 127, which made it the fourth worst performing trust in the country. On 4th June 2007, the Trust’s Clinical Quality and Effectiveness Group met and decided to take a number of actions, all focused not on reviewing or improving clinical quality or effectiveness, but on changing how activity was coded. They instructed coders not to use the codes that seemed to be contributing most to the high mortality score. A review of the case notes for patients who had died was conducted, not to establish if there had been any weaknesses in how they were treated, but to look for possible errors in the coding. Nine out of 14 deaths coded as ‘syncope’ were reviewed and the diagnosis changed in six of them. Six out of 11 deaths coded as ‘abdominal pain’ were reviewed and recoded. The West Midlands Strategic Health Authority was also concerned, especially since Mid Staffs was not the only hospital on its patch with a poor HSMR. They commissioned Richard Lilford and Mohammed Mohammed at Birmingham University to examine
whether the cause of the problem could be a flaw in how the HSMR was calculated.

**Problems with HSMRs**

Lilford and Mohammed had already published a series of criticisms of HSMRs.9,10 The HSMR is derived from HES data and has all the weaknesses that that implies. It ignores deaths occurring shortly after discharge and ignores deaths from the less common diagnoses. It is an unlikely measure of a hospital’s performance since, typically, 98% of inpatients survive their visit, so nearly all the data about the hospital’s activity are being ignored. Not only that but, of the 2% who do not survive, only a proportion will have died an ‘avoidable’ death, and it is only the avoidable deaths that are a measure of the quality of care. Simulations show that the HSMR is only a reliable measure of quality if the proportion of avoidable deaths is at least 15% of the total.11 In a typical hospital with, say, 10,000 admissions a year, you might expect 200 patients to die. Unless at least 30 of them are dying as a result of the hospital’s mismanagement, that mismanagement will not be picked up by the HSMR. For many this figure seemed ludicrously high.

Lilford and Mohammed have also argued that the HSMR is susceptible to bias and it does not generate random noise, but consistently favours some hospitals and penalizes others.12 Say two hospitals are equally good but hospital A has a higher proportion of emergency admissions than hospital B. One might expect that A, even if equally good, will have a higher death rate than B. Jarman and Aylin would say that they can still make a fair comparison because their measure takes the proportion of emergency admissions into account. Lilford and Mohammed term this the ‘case mix fallacy’. They say that of course emergency admissions are more dangerous than others, but the degree of danger will vary. For example, if hospital A is in an area where many patients are not registered with a GP, a high proportion of relatively healthy patients might be admitted through Accident and Emergency (A&E). The HSMR will treat these patients as at an equivalent risk compared with those admitted through A&Es elsewhere, wrongly giving the hospital a higher expected mortality and a low HSMR. Such biases are inevitable in an adjusted measure such as HSMR. Hospitals that are better at recording patients’ co-morbidities will also have a higher expected mortality, again making it easier to score well on the HSMR. Similar effects are found, for example, where patients are more easily discharged into a hospice, since the mortality rate used in the HSMR excludes those who die immediately after discharge.

This susceptibility to bias seems also to have been noticed by a competitor of DFI: CHKS, ‘a leading provider of healthcare intelligence and quality improvement services’. CHKS had advised Medway NHS Trust and had managed to lower their HSMR dramatically.13 Figure 1 shows how that change was achieved, by increasing Medway’s use of Z51.5, the ICD code for palliative care.

![Figure 1](http://www.curethenhs.co.uk/data-quality-and-clinical-coding/)  
*Figure 1*  
palliative care. The increase was, for a period, dramatic and had an impact on the HSMR. It is also worth noticing that the use of the code at the outset was clearly below the average and that it is now in line with normal practice. There is no evidence that coders at Mid Staffs were influenced by CHKS; however, the proportion of deaths coded as palliative care rose from 0% in the last quarter of 2007 to 34% in the third quarter of 2008.

The impact on the trust’s HSMR can be seen in Figure 2, falling below 100 for the first time since the 1990s. Francis found no evidence of a deliberate attempt to rig the results and considers other possible explanations for the change in coding practice: a new coding manager was in post, Government advice on the use of the code had changed. It is still striking that, in Jarman’s phrase, the hospital seems, in its coding practices at least, to have reinvented itself overnight as a specialist in terminal care. The timing is particularly suspicious: the changes coincided with the launch of a Healthcare Commission Inquiry into Mid Staffs in March 2008.8

Francis accepted that the trust managers were justified in considering the coding a possible source of the high HSMR, but they should nevertheless have checked that there was not a real threat to patients’ safety.

Jarman and Aylin have rebutted Lilford and Mohammed’s criticisms of their measure, arguing that it correlates with other indicators of the quality of care and that the impact of the biases identified by Lilford and Mohammed is, in practice, small.14–16 It is perhaps surprising that HSMR can detect failing hospitals, but perhaps that is because failing hospitals are worse than we expect. Between 2005 and 2008, the difference between the actual and expected number of deaths was 18% of the total, a higher figure than the 15% Lilford and Mohammed thought implausible.

Comparing mortality rates in paediatric cardiac surgery units

In the forward to his report, Francis refers to the Bristol Inquiry, suggesting that the necessity for his inquiry might be seen as the consequence of a failure to learn the lessons of the earlier one. Jarman has stated that it was the experience of the Bristol Inquiry that led him to campaign for the publication of hospital mortality rates. Yet the failings of the Mid Staffordshire NHS Trust were quite different from those of the Bristol Royal Infirmary. The focus of the inquiry into Mid Staffs has been less on the competence of the doctors and more on the quality of the nursing and on the failings of those whose duty it was to ensure that the nursing care was good.

After Bristol, Spiegelhalter, Aylin and Jarman all continued to look at the performance of centres doing paediatric cardiac surgery. Spiegelhalter published an analysis indicating an association between mortality rates and volume of surgery, although he declined to conclude that closing the centres with the smallest caseloads would save lives.17 Aylin, working with...
Jarman and others, published the data shown in Figure 3 in the BMJ, showing the performance of different units doing open heart surgery on patients under 1 year old during three periods in time. The analysis is based entirely on HES data. Each unit is shown as a red dot, the position of the dot is their best estimate of the mortality rate and the error bars show the 95% confidence interval. Centres are arranged in order of the number of operations performed, with smaller units to the right. The extent to which the performance at Bristol was ‘divergent’ during the period covered by the Inquiry is clear in the top figure. Oxford, the smallest of the centres, also seems to be significantly high; indeed doctors at Oxford reported Aylin to the General Medical Council for putting this figure into the public domain. Their case was dismissed and the unit is no longer performing open heart surgery on children under 1 year old.

In 2012 an NHS review recommended concentrating congenital cardiac surgery in seven centres. The

Figure 3 Mortality rates from the 11 different centres performing open heart operations on children aged under 1 year. Figure taken from Aylin et al. Published with permission
centres that were threatened with closure, including one at Leeds General Infirmary NHS Trust, campaigned against the decision and on 28 March 2013 Leeds succeeded in getting a judge to rule that the consultation process was flawed, granting the centre an apparent reprieve. The next day, however, Sir Bruce Keogh, the Medical Director of NHS England, announced that concerns about safety obliged him to suspend children’s heart surgery at the unit with immediate effect, a decision which was then reversed 11 days later, following the publication of a report from the National Institute for Cardiovascular Outcomes Research (NICOR). \(^20\)

NICOR’s analysis was based on a model called PRAIS that was developed to help surgical teams review their outcomes. \(^21\) The model is, like the expected mortality component of the HSMR, a risk model. Logistic regression applied to a 70% subset of 10 years’ worth of data from cardiac outcomes audits (26,447 episodes) was used to build a risk equation with terms for diagnosis, procedure, age, weight and co-morbidity. The model was validated on the remaining 30% of data. The model has been assessed in conjunction with a graphical device, the Variable Life-Adjusted Display (VLAD), by multidisciplinary clinical teams at mortality and morbidity meetings. \(^22\) VLAD charts show the difference between actual mortality and mortality predicted by PRAIS, not as a single ratio but a plot over time. The plots show a line rising with each successful operation and falling at each peri-operative death.

One of the difficulties with the use of mortality ratios to assess performance is in finding a presentation of the results that gives a suitable indication of the associated uncertainty. VLAD charts are one approach. Spiegelhalter has advocated the use of funnel plots, in which the estimated performance of the unit (i.e. the mortality ratio) is plotted against a measure of the precision associated with that estimate (i.e. some indication of the sample size on which it is calculated, such as the figure for expected deaths). \(^23\) Control limits plotted with these axes will be wider for the lower levels of precision and narrower at higher levels, hence the term funnel plot. There will, inevitably, be more imprecision over estimates of expected mortality in smaller units than in larger units, and therefore smaller units with problems will be more difficult to detect than larger units. These funnel plots avoid some of the difficulties associated with, for example, league tables of performance which
will tend to impose an ordering that the data may not support.

The NICOR report includes a table showing missing data for patient weight in data submitted by 11 centres in August 2012. Six centres have no missing data, in four less than 1.5% is missing, 34.7% was missing from Leeds’ submission. There is a clear implication that the decision to close the unit was based on incomplete data and that the decision was revised when the missing data were supplied. The authors observe that ‘the effectiveness of the data submission process could be considered as a measure of organizational culture and commitment to quality service delivery’. That observation notwithstanding, a funnel plot, Figure 4, of actual mortality compared with expected mortality calculated by PRAIS shows Leeds close to but not over an ‘Alert’ limit. The director of NICOR chose to resign rather than endorse the unit as safe—it is worth noting that the ‘Alert’ limit is fixed at an arbitrary threshold—but the report’s authors gave the unit a clean bill of health and it was reopened.

Conclusions

The monitoring of hospital death rates, both across hospitals and within units, remains a complex and evolving topic. The current government places a strong emphasis on transparency and argues that the publication of these data will drive up quality and enhance trust. The focus is, however, on identifying ‘outliers’. Some of the thinking here is influenced by work that Paul Aylin was commissioned to do by the inquiry into the career of Harold Shipman. Aylin applied statistical process control, an approach used in manufacturing to distinguish the rate of faults that occur when things are running as usual, from a rate that might be an early sign of malfunctioning. The metaphor of a single broken part in a system, with consequences of which one would want an early warning, seems to work in the Shipman case. It may not apply to hospitals. Lilford and Mohammed make the point that ‘the effectiveness of the data submission process could be considered as a measure of organizational culture and commitment to quality service delivery’. That observation notwithstanding, a funnel plot, Figure 4, of actual mortality compared with expected mortality calculated by PRAIS shows Leeds close to but not over an ‘Alert’ limit. The director of NICOR chose to resign rather than endorse the unit as safe—it is worth noting that the ‘Alert’ limit is fixed at an arbitrary threshold—but the report’s authors gave the unit a clean bill of health and it was reopened.

The NICOR report includes a table showing missing data for patient weight in data submitted by 11 centres in August 2012. Six centres have no missing data, in four less than 1.5% is missing, 34.7% was missing from Leeds’ submission. There is a clear implication that the decision to close the unit was based on incomplete data and that the decision was revised when the missing data were supplied. The authors observe that ‘the effectiveness of the data submission process could be considered as a measure of organizational culture and commitment to quality service delivery’. That observation notwithstanding, a funnel plot, Figure 4, of actual mortality compared with expected mortality calculated by PRAIS shows Leeds close to but not over an ‘Alert’ limit. The director of NICOR chose to resign rather than endorse the unit as safe—it is worth noting that the ‘Alert’ limit is fixed at an arbitrary threshold—but the report’s authors gave the unit a clean bill of health and it was reopened.

The Summary Hospital Mortality Indicator (SHMI) is a slightly more comprehensive measure than HSMR in that it includes deaths within 30 days of discharge (one area of progress in this field is in the linkage of different datasets so HES data can now be linked to the registers of deaths held by the Office of National Statistics) and is based not on the 50 most common diagnoses but on all diagnoses.26 It is in other respects simpler, using fewer variables to measure case-mix and taking no account of interactions between variables. Individual hospitals include their SHMI in a quality report which is published on the NHS Choices website. Data from the previous 12 months are published quarterly and hospitals are banded as being within, above or below control limits that are calculated as the 2 standard deviations from the mean, corresponding to a 95% control limit, applying a 10% trim from the top and bottom of all providers to allow for ‘over-dispersion’. Over-dispersion is a statistical concept reflecting that the variation between hospitals is greater than would be anticipated, so the effect is to inflate the control limits. It is worth noting that both the 95% control limit and the 10% dispersion are fixed entirely arbitrarily. In the most recent publication, the Health and Social Care Information Centre lists five trusts for which the SHMI is above the control limits and 11 for which it is below limits.27 These data are published alongside various quality indicators, including for example the proportion of patients coded as receiving palliative care. The fact that 11 trusts are ‘outliers’ in terms of having an unexpectedly low mortality rate has received rather little attention, but it suggests that perhaps there is weakness in an approach which focuses on comparing bad hospitals with average ones, since clearly even average hospitals could be improved.

The Care Quality Commission (CQC) now monitors 170 acute trusts and generates alerts for ‘excessive mortality’ in each of 610 procedure groups. The approach here is to compare actual mortality with expected mortality based on age, gender and time period and to look for persistently high score.28 There is an obvious advantage in a more fine-grained analysis than the whole-hospital measures; the problem is that in theory there could be 200000 figures tested for alerts each quarter. The question of where to set the correct threshold for alerts is therefore of paramount importance. It is possible to set the threshold and allow large numbers of false positives if the regimen for investigating them is appropriate. The CQC set a threshold which they describe as generating around 30 alerts a quarter.29 These are reviewed internally at the CQC and if necessary then pursued with the trust. Information about alerts is only published once the CQC is satisfied that the issue has been resolved.
A particular difficulty with this approach emerges when the performance of a failing hospital is quantified in terms of avoidable deaths. Subtracting the expected number of deaths from the actual number of death in a hospital over a period gives a measure of the ‘excess deaths’. There are sometimes characterized as ‘avoidable’ deaths. Of course, if all hospitals were equally effective, random variation would mean that ‘avoidable’ deaths would be detected in half of them. The estimate of ‘avoidable’ deaths that results from this calculation can be contrasted with a figure that is determined by getting experts to review the notes for a sample of patients who died in hospital and asking them to judge if the death could have been avoided. One study puts this latter figure at 5.2% of adult hospital deaths. Extrapolated across the NHS this would mean a total of 11,859 preventable deaths. Much larger figures, derived from HSMRs, have been widely reported in the UK media with the figure of 1200 avoidable deaths gaining widespread currency as a measure of the failures at Mid Staffs. In fact the HSMRs suggest that over the 3 years from 2006 to 2008, between 391 and 595 more patients died at Mid Staffs than would have been expected. The difficulties with the HSMR mean that we cannot be sure that real number falls within those limits. What can be said with certainty is that there was enough evidence of a problem for someone to have done more than change the way deaths were coded.

Acknowledgement
The author acknowledges that some of the material in this paper was originally published in the London Review of Books (11 April 2013), and thanks the editors for allowing its reproduction.

Conflict of interest: None declared.

References
10 Pitches DW, Mohammed MA, Lilford RJ. What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. BMC Health Serv Res 2007;7:91.
13 Hawkes N. How the message from mortality figures was missed at Mid Staffs. BMJ 2013;346:f562.


19 Dyer C. Consultation process on closing children’s cardiac surgery services at Brompton Hospital was fair, judges rule. BMJ 2012;344:e2896.


