Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma

Richard Lilford, Mohammed A Mohammed, David Spiegelhalter, Richard Thomson

The history of monitoring the outcomes of health care by external agencies can be traced to ancient times. However, the danger, now as then, is that in the search for improvement, comparative measures of mortality and morbidity are often overinterpreted, resulting in judgments about the underlying quality of care. Such judgments can translate into performance management strategies in the form of capricious sanctions (such as star ratings) and unjustified rewards (such as special freedoms or financial allocations). The resulting risk of stigmatising an entire institution injects huge tensions into health-care organisations and can divert attention from genuine improvement towards superficial improvement or even gaming behaviour (ie, manipulating the system). These dangers apply particularly to measures of outcome and throughput. We argue that comparative outcome data (league tables) should not be used by external agents to make judgments about quality of hospital care. Although they might provide a reasonable measure of quality in some high-risk surgical situations, they have little validity in acute medical settings. Their use to support a system of reward and punishment is unfair and, unsurprisingly, often resisted by clinicians and managers. We argue further that although outcome data are useful for research and monitoring trends within an organisation, those who wish to improve care for patients and not penalise doctors and managers, should concentrate on direct measurement of adherence to clinical and managerial standards.

Concerns about quality and safety of care combined with notions of clinical governance, professional accountability, public disclosure, and reactions to prominent health care tragedies, such as the Shipman murders and paediatric heart surgery in Bristol, UK, have made performance monitoring of health care organisations in general and intensive care units (ICUs) in particular an integral part of modern health care. However, the roots of monitoring health care organisations can be traced to much earlier times with notable names such as Florence Nightingale, Ernest Codman, and Lord Moymihan amongst others. It is hard to argue against making data publicly available so that people can make of them what they will. Newspapers might want to publish them, researchers and managers investigate them, and prospective patients could use them to select their caregiver. But these data can also be used by agencies outside the individual health care organisations, whose business is “performance management”; issuing reward and punishment on the basis of judgments about quality of care based on comparative data (panel). These agencies might be part of, or responsible to, the central administration of a multi-provider organisation, such as the National Health Service in the UK or the Department of Veterans Affairs in the USA, or they might be purchasers of health care such as Medicare or Medicaid in the USA. It is with the use of data to make external judgments and to apply rewards (eg, financial gain or special freedoms) or sanction (eg, star ratings) that this article is concerned.

We start from a moral premise and a managerial principle. The moral premise is that if external agencies use aggregated comparative data for judgment, this action should be fair—ie, the data truly reflect underlying differences in quality. The management principle builds on this since, as we shall show, unfair comparisons provoke inappropriate management responses.

Outcome data

During the 1980s and 1990s, the concept of outcome measurement became popular, and it was based on the premise that outcome is the ultimate measure of quality of care. This notion can be traced back to Ernest Codman’s end-results idea. Outcome data can be patient-rated (satisfaction and quality of life) or recorded by an external party (mortality and morbidity). For now, we use outcome as short-hand for observed mortality and morbidity. Use of outcomes to compare quality of care implies that the variation due to other causes can be accounted for, such that any residual variation truly indicates quality of care variation. Is this quixotic?

Outcomes—what do they tell us?

Outcomes are influenced by definitions, data quality, patient case-mix, clinical judgment of care and chance.

Search strategy and selection criteria

We started our search for papers dealing with correlating quality of clinical care with outcomes, with the classic 1997 contribution from lezzoni. We identified cited papers dealing with the relation between quality and outcome, and obtained relevant MESH headings which we then used as the basis for a systematic literature search in MEDLINE. From the initial yield of more than 5000 papers, we identified those that attempted to quantify the association between quality of care and outcome, which we are now reviewing in detail.

Lancet 2004; 363: 1147–54

Department of Public Health and Epidemiology, University of Birmingham, UK (Prof R Lilford MD); Department of Public Health and Epidemiology, University of Birmingham (M A Mohammed PhD);
MRC Biostatistics Unit, Institute of Public Health, Cambridge (D Spiegelhalter PhD); School of Population and Health Sciences (Epidemiology and Public Health), University of Newcastle (Prof R Thomson MD)

Correspondence to: Richard Lilford
(e-mail: R.J.Lilford@Bham.ac.uk)
Performance measures and indicators

The distinction between a measure of quality and an indicator of quality is important. Generally speaking we have very few real measures of quality. For example, post operative length of stay is a measure of the patient’s hospital stay, but only an indicator of quality—eg, a patient’s long stay might represent postoperative complications or poor discharge arrangements. Thus, the term indicator is preferable.

Performance monitoring

Regular review of performance by use of any combination of structure, process, or outcome data.

Performance management

External judgment of the quality of care based on performance monitoring data followed by a system of reward and punishment.

Process of continual improvement

Application of the scientific method to deliver improvement based essentially on an understanding of variation and the plan-do-study-act cycle, akin to the hypothesis generation and testing cycle of the scientific method.

Safety, medical error, and quality of care

Safety is defined as the absence of clinical error, which can be classified as errors of commission (unintentionally doing the wrong thing) or omission (unintentionally not doing the right thing). So although performance monitoring might focus on safety and error, it is not possible to disentangle this from quality of care. We therefore use the term quality of care in its broader sense which includes medical error and safety.

Definitions and data quality

Outcomes are based on measurement of numerators and denominators. A frequently used numerator in performance monitoring of acute care is death, but even with this numerator systematic differences arise in definitions—for example, by classifying dead on arrival at the hospital as a community death or instituting early discharge policies to shift the mortality burden. Morbidity numerators, such as wound infection, anastomotic leak, and postpartum haemorrhage, are notoriously subject to systematic differences in definition. Denominators can also be counted in different ways. Whereas fractured neck of femur or birth are clear cut entities, myocardial infarction, heart failure, stroke, and infertility are harder to classify, and institutions can vary in their definitions and in collection and coding of data. Ranking of hospitals depends upon the data source used, varying markedly, for example, if the same outcome data were obtained from case records or from administrative data sets.

Case-mix

Fair comparison of providers, including ICUs, is believed to need consideration of differences in case-mix, and case-mix adjusted mortality comparison of hospitals with ICUs is widely advocated and practised. Nonetheless, case-mix adjustment can lead to the erroneous conclusion that an unbiased comparison between providers then follows. We term this the case-mix fallacy.

Clinical care

Clinical quality of care factors are essentially the processes by which resources are used by clinicians and managers. Some aspects of quality of care are of proven effectiveness but some are tacit—such factors (eg, surgical skills and team working) have so far proved difficult to measure directly, although attempts have been made. We discuss later the notion that clinical quality is itself largely a function of systems within which care is embedded.

Sources of variability

The informal equation which links the total variance in outcome—between health-care providers to the components of variance (V) can be written as:

\[ V(\text{outcome}) = V(\text{definitions/data quality}) + V(\text{case-mix}) + V(\text{clinical quality of care}) + V(\text{chance}) \]

Case-mix adjustment is by far the most widely used method to analyse variance in outcome. Much has been written about different statistical methods to produce case-mix adjusted outcomes. Different case-mix adjustment methods identify different provider units as low or high performing. However, with the exception of chance, every term in the equation has components that can be measured and those that cannot.

Therefore, even if an agreed risk-adjustment method could be derived, outcomes could still vary systematically between providers because we can never be sure that risk adjustment is not hampered by unmeasured prognostic factors, such as how well patients were cared for before admission, different attitudes to “do not resuscitate” orders, or systematic differences in case-mix not recorded in the database. Making judgments about quality of care on the basis of risk-adjusted comparisons cannot guarantee that like is being compared with like. Also, the sensitivity of an institution’s position in league tables to the method of risk adjustment used suggests that comparisons of outcomes are unlikely to tell us much about the quality of care.

Correlating quality of clinical care with outcomes

In several studies, researchers have combined case-mix adjusted outcomes with measures of clinical quality of care, obtained at the same time from the same institutions. Many of these investigators found no association between quality and outcome. For example, Park and colleagues recorded no correlation between outcome and quality of care for congestive heart failure or pneumonia. Best and Cowper noted no significant associations between outcome and quality of acute medical care in hospitals with high mortality. Jenks and co-workers did not find any correlation between quality of care and the outcome of myocardial infarction. However, others have identified a weak association, for myocardial infarction, a range of acute medical conditions, hip fracture, and stroke. Hannan and colleagues noted more striking quality of care problems in deaths after coronary artery bypass grafting (CABG) in high mortality rather than low mortality hospitals. However, restricting scrutiny to deaths is potentially confounded by differences in pre-operative risk between patients who die at different hospitals.

Thomas and Hofer reviewed 18 articles about the relation between outcome and clinical process and quality. They concluded that overall quality of care has some correlation with outcome but that it is weak, so that most hospitals in the highest 5% for mortality (outliers) will not be among the 5% providing the poorest quality of care and most poor quality care will not reside among the outliers—at least for acute medical care. CABG surgery is a possible exception, although this topic remains highly controversial.
We now have an empirical answer to our question—is it unrealistic to use outcome data to compare quality with the confidence necessary for performance management? The answer is yes; outcome is neither a sensitive nor a specific marker for quality of care. Case-mix adjustment, it turns out, does not allow the residual to be attributed to quality of care with any reliability and so cannot be used to support fair comparisons between providers (with the possible exception of a few surgical procedures). As we will argue below, measurement of outcomes for research purposes is useful to help organisations detect trends and to spot extreme outliers but league tables of outcomes are not a valid instrument for day-to-day performance management by external agencies. That is to say, sanction and reward should not be applied to the “worst” 5% of providers on outcome, because these will not be the 5% with the worst quality.

If not outcomes then what?

We need performance measures which better reflect the quality of care. We will come later to the distinction between performance management (systems of punishment or reward) and continual non-judgment improvement.

Figure 1 shows a conceptual map of factors that could affect the final quality and quantity of care. The map depicts a causal chain starting from the structures in which management processes are nested. Structural factors are those that cannot easily be affected at the organisational level because they depend on release of substantial resources or changes in policy. Institutional process factors include those that are the responsibility of local managers and can generally be affected with modest re-allocation of resources (eg, human resource policies). Institutional factors affect intervening variables, such as how hard or carefully clinicians work and how knowledgeable they are, which in turn affect throughput and clinical processes. Can we measure any of these factors and, if so, what do they mean?

Structural factors and institutional processes

Several measurable structural and institutional factors are associated with clinical outcomes. Although few if any have been tested in prospective trials, the association is quite strong in some cases, and cause and effect conclusions are plausible, since one structural or institutional process can affect many clinical processes. For example, the quality of care at top and bottom ranking Veterans Affairs hospitals in the USA was assessed with case-note review50 and structured site visits.47 Differences in structural factors (eg, availability of equipment and staffing levels) correlated with outcome.50 Method of reimbursement51,52 and teaching hospital status53 are also associated with outcome. Several institutional management processes are associated with improved outcomes in ICUs. These included daily patients’ rounds,54 “closed” ICU care (in which patients are managed primarily by a full-time intensivist54), and good communication between clinical staff.55 Many human management processes seem to correlate with outcome.48–51,57–61

All the above correlations are associations for which inferences about cause and effect should be made with caution, but pending experimental proof, which is woefully absent in organisational research,47 incorporation of these findings into practice is reasonable. The implications for practice belong to two rather different categories. Structural changes need large-scale investment—improving the nurse to patient ratio, for example. Institutional changes involve better management of existing resources—ensuring that all staff are appraised or improving communication within teams, for example.

Although third party funding agencies, such as Medicaid, might not want to draw a distinction between these two types of factor when placing contracts—why should they care if some hospitals have a natural advantage—the situation is different within individual systems such as the NHS in the UK or Veterans Affairs in the USA.

The distinction between factors that are in the control of an organisation and those that are not is crucial within a single system. Penalising providers for deficiencies for which central management is responsible is as perverse as the Pharaoh’s punishment of the Hebrew slaves while failing to provide the raw materials to make bricks. Indeed, it is an example of the downward cascade of
blame in command economies such as the NHS; policy makers shift responsibility to managers, and managers to clinicians. Head office should base any performance management on institutional processes, which local management can influence. Ultimately, however, it is clinical care that counts. Why not measure clinical processes directly in quality improvement efforts?

**Clinical process measures**

Measurement of clinical processes offers advantages over outcome based monitoring as a practical instrument to stimulate change. Clinical process measures should be based on agreed criteria, supported by evidence, or logic, and include actions such as appropriate use of β blockers after acute myocardial infarction, the use of lower tidal volume in acute respiratory distress syndrome, and avoiding delay in the use of antibiotics in pneumonia. Clinical process measures guide efforts to improve performance because they are a direct measure of performance based on adherence to established clinical standards. Monitoring clinical process has several advantages over outcome monitoring: (1) it focuses on violation of agreed and evidence or logic based standards, so that failure is failure, not an indirect and inaccurate indicator thereof; (2) measurement can be made close to the point of delivery of care, overcoming the delay between intervention and outcome; (3) the target for action is inherent in the measurement made, thereby avoiding institutional stigma; (4) it can be applied to all institutions, not just the “worst” 1%, 2%, or 5%, and therefore offers the hope of improving the average quality of care, yielding far bigger gains to the public health (figure 2).

Clinical process monitoring needs access to information which, although expensive, is likely to be much more cost effective than outcome monitoring. Mant and Hicks estimated that plausible differences in quality of clinical care might result in a 10% difference in mortality across hospitals and that it would be necessary to assess 3619 patients from each hospital to provide a reasonable chance of detecting this. However, only 48 cases would need to be assessed in each hospital to detect the corresponding difference in adherence to quality standards. Nonetheless, process-based monitoring is subject to potential bias due to the fact that the opportunity for error varies by case-mix; sicker patients need more care, which gives greater opportunity for errors of commission and omission. We therefore advocate measuring process of care as a function of “opportunity for violation of the standards”. Some have argued that rather than assess all cases treated, source materials need be evaluated only in those cases in which there is reason to suspect an error or poor care—ie, triggers such as re-admissions or prolonged stay. This type of assessment might be beneficial for monitoring within institutions, but it is suspect if used to make comparisons or inferences between institutions, because triggers are prone to case-mix bias just as outcome measures are. Furthermore, triggers (apart from those relating to medication error) correlate poorly with the overall incidence of violations in quality, although this is a topic of further investigation sponsored by the NHS Patient Safety Research Programme.

**Throughout**

Some process measures are based on management data rather than adherence to clinical standards. These measures include waiting lists, ambulance response times, and delays in accident and emergency departments. Such performance data are potentially useful for quality improvement but when used for performance management they often lead to a focus on changing the numbers rather than genuinely improving the systems, just as quotas led Soviet farmers to play the system (eg, by deflating declared harvest size in order to produce apparent improvement in subsequent years).

**Patient-rated outcome**

We believe that organisations should measure and respond to the opinions of their service users. We also believe that it is reasonable for external organisations (eg, head office, service commissioners) to ensure that service providers do that. However, it is wrong to compare organisations (for performance management purposes) on the basis of differences in patients’ satisfaction or quality of life. Patient-rated outcomes vary by many features such as age, wealth, and ethnic background. So why not measure these confounders and adjust for them statistically to compare and penalise or reward institutions? The reason, yet again, lies in case-mix fallacy.

**The role of performance monitoring—to judge or not to judge?**

Imagine that you are managing a hospital ranking on the 98th centile for case-mix adjusted hospital mortality, and that this placement has contributed to you achieving only one star in a three star grading. You would not know whether your poor showing was due to: differences in definitions and data quality; chance, although the effects of this are quantifiable statistically; case-mix differences for which risk adjustment was inadequate; structural factors affecting clinical processes; or institutional management factors affecting clinical processes; clinical skill and diligence unaffected by structural and institutional constraints; or tacit or currently unknown clinical factors.

The appropriate clinical and managerial response should turn on which of these is causal. Chance and case-mix would need no response contingent on the ranking. Violation of specific clinical standards would need immediate, targeted, correction. Differences in outcome caused by structural institutional factors would need policy action by central administration. Institutional management factors would need changes in management of the hospital. The crux of the performance management problem, however, is that you cannot know which of these factors is operating when outcomes differ. Thus, we should not be surprised if outcome-based performance management causes tension. Blame is attributed without specification of what the problem is and where it lies. It is in this respect that health care differs so radically from many industries—showing people...
that their outcomes are worse than others does not automatically tell them what to do to improve or even whether they have a greater need than others to improve. Small wonder then, that dysfunctional behaviour can result from the ensuing fear and distrust, and impede quality improvement efforts. Examples of dysfunctional behaviour are documented in a review monitoring performance of coronary artery bypass graft surgery.26

Such behaviour can be categorised as gaming (or manipulating the system)—gaming with the data, gaming with the patients, or gaming with the process of care. Green and Wintfeld27 noted that 41% of New York State’s (USA) reduction in risk-adjusted mortality from CABG could be accounted for by data gaming—an increase in risk factor coding or a change in definitions.27 Burack and colleagues28 surveyed 104 cardiac surgeons in New York and found that high-risk CABG patients were more likely to be denied treatment compared with similar high risk patients with aortic dissection, because the latter procedure was not subject to external performance monitoring. Carey and co-workers29 noted how some surgeons changed the operative class of CABG–only patients to CABG–mitral valve repair by adding a few commissural stitches. By contrast, a judgment-free programme, the Veterans Affairs National Surgical Risk Improvement programme, reported a 27% reduction in mortality and 45% reduction in morbidity over 5 years with no change in the patients’ risk-profiles.30 They tried to ensure that their system was neither penalising nor stigmatising. They helped hospitals by providing a central outcomes service with anonymous feedback and by encouraging local innovation.

There is also evidence of dysfunctional behaviour associated with measures of throughput.31 An example is the use of the UK Patient’s Charter indicator for the proportion of patients seen within 5 minutes of presentation at accident and emergency departments. Hospitals performance was judged on this indicator. The effect was to stimulate the wide appointment of the pejoratively termed “hellos nurse”32 thereby ensuring the target was met without necessarily improving triage.33 Another example34 concerns emergency departments, anxious to keep their accident and emergency department waiting times low, who refuse to allow patients brought in by ambulance to cross the hospital threshold until they can be seen—patients are deposited in inflatable tents in a performance target free zone. In the UK, we have seen clinical services interrupted, figures distorted, and managers falling on their swords in response to spurious process data35 or “bean counting”.

We conclude that the use of outcome data and throughput to judge quality of care should be diligently avoided. We have shown that when such measures are used to judge quality of care, they often result in several predictable reactions ranging from resistance to gaming. This response was predicted by Deming,36 a professor of statistics and an international management consultant, several decades ago. He argued that management must drive out fear, eliminate exhortation to meet targets and improve processes—eg, to improve throughputs. Wherever we set this threshold, provided that it is not zero or infinity, we need to be alert to possibilities of misclassification and be aware that the threshold may need to vary according to context. Furthermore, methods for identification of the underlying cause for an outlier organisation need to be predicated on probable previous causes that place the individual clinicians at the last part of the investigation process91 (figure 3). According to Deming,36 the underlying cause is in most instances attributable to the system and not the individual. And even when responsibility rests heavily on one clinician, the outcome analyses have shown they could have been spotted earlier.39–41 the implications for management were unclear until the reason for the unusual results had been investigated. Subsequent investigations uncovered a murderer in one case, poor systems of care in the second case,39 and a cardiac surgeon operating with an undiagnosed brain tumour in the third case. Even in these most extreme situations we are unable to reliably use outcome data to judge where the quality of care was deficient.

So, to move away from judgment and towards improvement, we suggest that the development and implementation of judgment-free, scientifically rigorous methods37 is necessary, whereby clinical services monitor their own performance (process and outcome), compare themselves with others or their own past performance as appropriate, and take whatever action seems necessary. Additional action from the centre should only be taken when an extreme threshold is passed—external agencies might prefer a lower threshold than clinicians and local managers, although a conventionally used threshold in continual improvement methods is three sigma (SD) from the mean.42 Wherever we set this threshold, provided that it is not zero or infinity, we need to be alert to possibilities of misclassification and be aware that the threshold may need to vary according to context. Furthermore, methods for identification of the underlying cause for an outlier organisation need to be predicated on probable previous causes that place the individual clinicians at the last part of the investigation process91 (figure 3). According to Deming,36 the underlying cause is in most instances attributable to the system and not the individual. And even when responsibility rests heavily on one clinician, this can be remedied humanely43 without a witch-hunt. Poor performance is not a sign of moral turpitude unless it is wilful or deliberately concealed.
However, focus on outliers is only one aspect of improvement. Improving clinical and managerial processes in the remaining organisations can achieve much more health gain by shifting the mean (figure 2). The actions required to pull back an outlier as opposed to shifting the mean (figure 3) are fundamentally different. The first needs investigation of a probable special cause, whereas the second action needs focus on improvement by means of a wide array of process-improving tools such as the plan-do-study-act cycle.

For these reasons, hospitals should monitor their own data—institutional processes, throughput, clinical processes, outcome (especially for trends over time), and above all clinical process.

We have some recommendations for external agencies (purchasers, central administration, and their agents). We recommend that they encourage providers organisations to develop institutional processes to deliver continual improvement (whether they have a plan-do-study-act framework in each clinical service, whether service users are involved; or if human resources policies such as appraisal have been instituted). In this sense, the centre acts as a management consultant, not police officer. We suggest that they ask organisations to show that they have mechanisms in place to monitor trends in both outcome (eg, a sudden change in wound infection which might be traceable to a change in theatre cleaning policy), clinical processes (eg, fractured hips operated on within 24 h), and throughput (waiting times in casualty). What the external agency is doing here is not looking for bad apples, but expecting the organisation to have processes in place for self-monitoring. Individual organisations might want to compare themselves with others, and the head office can provide the means to do so, as done for example by the Veterans Affairs.

Central organisations could also institute national criterion based audits—ie, comparisons of adherence to established clinical standards known to enhance safety and improve outcomes, perhaps arranged according to specialty on a cycle (eg, Royal College of Physicians audit of Myocardial Infarction) or on randomly selected cases of all sorts (eg, Medicaid). National outcome data could be used to detect extreme outliers—beyond three sigma—as a precursor to non-judgmental investigation. Research across organisations should be sponsored to explain the causes of differences in outcome. Such studies may be retrospective (eg, natural experiments) or prospective (eg, cluster trials) but should not be based on single examples; so-called “benchmarking”. Rather, many instances (eg, lots of hospitals) should be examined, and repeated measurements (eg, time series) used to keep risks of selection bias and regression to the mean to a minimum.

Finally, the agencies should facilitate the development and dissemination of a database for best practice and improvement based on the results for primary and secondary research.

These recommendations can be implemented with little recourse to reward and sanction tactics of performance management which, in principle, seem to have changed little since King Hammurabi (1750 BC). He decreed that surgeons would be rewarded with ten shekels of silver for a good outcome, or be punished, after some allowance for case-mix, by having their fingers chopped off for a bad outcome. Performance monitoring by outcome and throughput simply substitutes the collective psychological pain of stigmatisation for physical pain, and we need to move on to methods which are both more effective and more humane.

Conflict of interest statement
R Thomson is Director of the UK Quality Indicator Project, a non-judgmental quality indicator system, and both he and D Spiegelhalter are expert advisers to the Commission for Health Improvement’s Performance Assessment Sub-Committee. The other authors have no conflicts of interest.

Acknowledgments
We acknowledge our colleague David Pitches who is leading author on the systematic review mentioned in the search strategy panel.

Reference
22 Julious SA, Mullee MA. Crude rates of outcome.

21 Hartz AJ, Kuhn EM. Comparing hospitals that perform coronary surgery.


17 Daley J, Henderson WG, Khuri SF. Risk-adjusted surgical outcomes.


INPATIENT SAFETY III


