QUALITY IMPROVEMENT RESEARCH

Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls

A E Powell, H T O Davies, R G Thomson

Qual Saf Health Care 2003;**12**:122–128

Measuring the quality of health care has become a major concern for funders and providers of health services in recent decades. One of the ways in which quality of care is currently assessed is by taking routinely collected data and analysing them quantitatively. The use of routine data has many advantages but there are also some important pitfalls. Collating numerical data in this way means that comparisons can be made—whether over time, with benchmarks, or with other healthcare providers (at individual or institutional levels of aggregation). Inevitably, such comparisons reveal variations. The natural inclination is then to assume that such variations imply rankings: that the measures reflect quality and that variations in the measures reflect variations in quality. This paper identifies reasons why these assumptions need to be applied with care, and illustrates the pitfalls with examples from recent empirical work. It is intended to guide not only those who wish to interpret comparative quality data, but also those who wish to develop systems for such analyses themselves.

See end of article for authors' affiliations

Correspondence to:
Professor H T O Davies,
Department of
Management, University of
St Andrews, St Andrews,
Fife KY16 9AL, UK;
hd@st-and.ac.uk

Accepted for publication
9 November 2002

The pressures facing healthcare funders and providers are mounting. Not only is health care itself becoming more complex, with an increasing range of new treatment options and competing approaches to the delivery of care, but care must also be delivered in a context of cost constraints, increasing consumer demands, and a greater focus on accountability. Against this background there has been an explosion in official and unofficial schemes aiming to use routine data to compare performance between healthcare providers.[1 2] Such data can help highlight any problem areas in clinical performance, inform or drive quality improvement activities, prompt reflections on clinical practice, and identify important issues for further research. These data thus have a wide range of potential uses and are of interest to a wide range of stakeholders—researchers, practitioners, managers, purchasers, policy makers, patients, and carers.

The logic underlying an analysis of routine comparative data is that it is possible to make attributions of causality between the services provided and the observed quality measures—that is, that high *measured* performance reflects good *actual* performance—and, conversely, that low measured performance reflects poor actual performance. For example, if one hospital has markedly better survival rates 30 days after myocardial infarction than another, then one conclusion from the data might be that the higher survival rates result from higher quality care at the first hospital. But how well founded is this conclusion? What are the difficulties that arise in developing schemes that can draw robust conclusions from routinely collected comparative data? This paper will consider a range of potential pitfalls. Using examples from recent studies of quality variations we explore the strengths and weaknesses of using routine comparative data to draw conclusions about quality of health care.

USING ROUTINE DATA IN HEALTH CARE

The use of routine data for quality assessment purposes is attractive to researchers, health professionals, managers, and policy makers. Box 1 summarises the reasons: the data are readily available, they are a potentially rich source of information about large numbers of patients, and using existing data is less demanding (and has fewer ethical constraints) than planning, funding and executing long term experimental studies.

What routine data have in common is that they are often collected for other purposes and are observational rather than experimental. Examples of routine data collected in the NHS, to take just one healthcare system, include perioperative deaths, hospital bed occupancy rates, use of contraceptive services in general practice, cervical screening, and vaccination rates (see, for example, the latest government collated data at http://www.doh.gov.uk/performancerating/2002/index.html). In the US, routine data collection includes state specific initiatives—for example, the California Hospital Outcomes Project reports public data on a range of indicators including 30 day mortality after myocardial infarction[3 4]—and national programmes—for example, the national register for myocardial infarction supported by Genentech.[5] In addition, the Joint Commission on Accreditation of Healthcare Organisations (JCAHO) requires hospitals to submit clinical performance data on six measures,[6] and extensive measures are collected to help compare health plans.[7] Indicators derived from routine data may cover processes of care (for example, treatments given; length of hospital stay), "true" outcomes (for example, mortality rates) or "proxy" outcomes (for example, physiological measures such as blood pressure or weight gain). In quality assessment terms, process measures—which assess, for example, what was done and when (how quickly or how often)—may be most illuminating

Box 1 Reasons for use of routine data for quality assessment purposes

- The data are readily available in many healthcare settings (although accuracy and completeness may vary).
- They can be used retrospectively whereas experimental designs by definition have to be set up prospectively; thus data for large time periods can be gathered more quickly than in a new study.
- Because the data are in many cases already being collected for other purposes, the costs of setting up data collection and retrieval systems are likely to be much lower.
- The data are a rich source of information about large numbers of patients with different conditions across diverse geographical and healthcare settings.
- Ethical and consent issues applying to routine data are less problematic than those which apply to data gathering primarily for research purposes.

if there is definite evidence for providing a particular drug treatment or intervention.[8] Outcome measures may be more useful if there is a clear temporal and causal link between the care given and the outcome achieved, and if there is consensus about the value to the patient and/or the service of the outcome studied.[9] In any case, routine data impose certain interpretation limitations on both process and outcome measures, although these limitations may operate in different ways.[10]

ISSUES IN INTERPRETING ROUTINE DATA TO ASSESS QUALITY OF CARE

Whether attempting to gain understanding from published comparisons of performance or designing new schemes to analyse routine data to allow such comparisons, we need to understand how the ways in which data are collected may impact on the interpretations that are possible. A clear understanding of the potential pitfalls arising may allow some of these to be anticipated and avoided during system design, or can temper the conclusions drawn from established schemes. Four main issues affect the interpretation of comparative routine data:

- Measurement properties
- Controlling for case mix and other relevant factors
- Coping with chance variability
- Data quality

Measurement properties

The development and validation of indicators is dealt with in a separate paper in this series.[11] Two key measurement concerns apply when routine data are used for quality purposes: (1) the validity and reliability of the outcome measures themselves; and (2) the risk of conflicting findings when different measures are used to assess the same organisations.

Validity and reliability

Routine data provide a given set of variables from which quality measures can be selected or derived. Yet poor validity and/or reliability of the measures can undermine the conclusions drawn. Common threats to validity and reliability may arise from many sources—for example, from the fact that such data are usually collected unblinded. When outcome assessors are aware of previous treatment histories there is empirical evidence that this may affect the judgements reached.[12] Furthermore, when providers are aware that they will be judged on the data, other incentives may come into play leading to concerns about "gaming" with data.[13]

Inappropriate data sources may add to measurement concerns, either overestimating or underestimating services provided. Hospital discharge forms, for example, which are produced for billing and other administrative purposes, may lack the important clinical details needed for quality assessment.[14 15] In one study of medical and surgical hospital discharges, a tool commonly used in the US to analyse computerised hospital discharge abstracts (the Complications Screening Program) failed to pick up instances of substandard care that were evident when the same patients' medical records were reviewed according to explicit criteria.[16] Another study in primary care which set out to determine the optimal method of measuring the delivery of outpatient services found that medical records had low sensitivity for measuring health counselling but moderate sensitivity for some other aspects including laboratory testing and immunisation.[17] In contrast, self-completed patient exit questionnaires had mixed sensitivity for laboratory testing but moderate to high sensitivity for health counselling and immunisation.

Even apparently "hard" and valid end points like death can be problematic. A study of routine data on in-hospital deaths[18] found that this measure gave an incomplete reflection of mortality within 30 days of admission for surgery, and that the measure was less valid in more recent years than historically. "In-hospital deaths" are usually defined as deaths in the admission in which surgery was performed, but counting only these deaths excludes those that occur elsewhere—whether at home or in other hospitals after transfer—but which may nonetheless be related to the operative care. Since these statistics were first collected in the 1960s and 1970s, shorter hospital stays and an increased tendency to transfer patients between hospitals for specialist care mean that a greater proportion of deaths within 30 days of admission for surgery will be missed if the original definition is used.

Additional problems arise because some clinical conditions, particularly chronic problems, may require a range of interventions provided by different health professionals, both in hospital and in the community, and thus may not be amenable to a single quality measure. For example, the US Health Plan Employer Data and Information Set (HEDIS), which is used across a range of services, was found to have low validity as a tool by which to assess behavioural healthcare quality after hospital admission for major affective disorder.[19]

The validity of quality measures derived from routine data may be further undermined by changes in reporting practices over time. In a study of emergency admissions in one health authority from 1989/90 to 1997/8 an apparent increase in emergency activity was not matched by an increase in the number of admissions or by the increase in the number of patients each year.[20] What appeared to be a rise in emergency admissions turned out to be mainly due to increased reporting of internal transfers after admission.

The validity of routine data may also be compromised by differences in data gathering practice between providers. A study of nosocomial (hospital acquired) infection[21] suggests that it is difficult to make meaningful comparisons between hospitals in relation to reported infection rates as the infection surveillance practices vary so much. The study investigated whether there was a relationship between surveillance practices and reported nosocomial infection rates but concluded that any such relationship was not systematic: "*there appears to be a serious issue regarding the equivalency of the data collection processes*".[21]

The overriding message from these diverse examples is the need to establish beforehand the validity, reliability, sensitivity, and other metric properties of any proposed measures. Wherever possible, indicators drawn from routine data should be tested against freshly gathered data using "gold standard" measures with well established metric properties.

Conflicting findings

Routine data sets may provide a wide variety of potential measures on which to assess quality of care—for example, quality of surgical services may be assessed using early or longer term mortality, duration of ICU stay, complications, or infection rates. Yet studies using routine data suggest that there may be little correlation between different outcomes, with the effect that the same institution may look good on some indicators and poor on others[22]—for example, there may be little or no correlation between hospital mortality rates and complication rates.[23] This means that the ranking of institutions and the selection of those requiring further quality review may largely depend on the specific measures chosen for review. When data are being used for exploration and insight, such conflicting messages may simply offer a rich source of avenues for further investigation. However, if the intention is to seek to make definitive judgements on performance, then the potential for conflicting findings suggests the need to make decisions on primary end points at the design stage (much as is now commonplace in the design of prospective randomised controlled trials).

Controlling for case mix and other relevant factors

Even when concerns over validity and reliability have been satisfied, there is a further threat to the meaningful interpretation of routine data. Performance comparisons between healthcare providers need to take into account whether the measures being compared derive from similar patient groups: clinical and other characteristics of the patients treated are likely to affect both the demands placed on the service and, in particular, the outcomes from treatment. In observational studies these case mix differences cannot be controlled for at the outset as they are through randomisation in prospective trials, so retrospective risk adjustment is required instead. Such systems usually rely on developing some kind of scoring system that encapsulates the level of risk the patient faces, irrespective of any care delivered. Clearly, accounting for these factors, which are outside the control of those providing the care, is essential before any comparison of the outcome of care is possible.

Various patient characteristics have been recognised as important in increasing risk, such as age, disease severity, co-morbidities, and past medical history. Scoring systems are designed to quantify a number of discrete but interrelated patient characteristics and reduce these to a single value reflecting the overall severity of the condition or risk that the patient faces. For each scoring system the association between the independent variables (patient characteristics) and the dependent variable (the outcome of interest, often death) is described in the form of a mathematical model known as a multiple logistic regression model. The mathematical model describes the strength of the association of each of the different independent variables with the dependent variable, while allowing for the effect of all the other independent variables in the same model. The weights or coefficients associated with the independent variables in the model are derived from analysis of large databases of past patients containing information on both the patient factors required for the scoring system and the outcomes of interest. These models must then be calibrated before being assessed for their sensitivity and specificity—most usefully, this testing procedure should be carried out on new cohorts of patients. Needless to say, developing and testing such scoring schemes is highly technical and complex, and requires substantial statistical expertise. Developing robust schemes thus introduces a wide range of challenges which are reviewed in detail elsewhere.[14 23]

Sophisticated risk adjustment models take a long time to create, test, and implement. It is necessary to know which aspects of case mix have any bearing on outcomes in order to know whether to adjust for these in any given patient group and to what extent, either singly or in combination with other factors. Further, the usefulness of the risk adjustment model is only as good as the underlying assumptions on which it is based, which means that there has to be a priori knowledge that a particular factor is relevant. For many diagnoses the necessary information is not yet available to create robust risk adjustment models.

Even with a reasonably robust risk adjustment model, the data demands of risk adjustment pose their own challenges which may increase as the complexity of the model increases. Sophisticated risk adjustment requires detailed information about which patients had these characteristics in the first place. This information is rarely routinely available—either because at the time the data were collected this particular characteristic was not thought to be relevant, or because the data set is incomplete or inaccurate in certain aspects. The concerns over validity and reliability of measures explored earlier apply as much, if not more, to the variables collected for risk adjustment as they do to the outcome measures themselves.

A further difficulty arises when different risk adjustment schemes lead to different rankings of performance. To take one example, risk adjusted mortality rates may not be a good measure of hospital quality of care as severity can be defined in quite different ways and different severity adjustment schemes may lead to different judgements.[23] This is well illustrated in a study which showed that the same patients were assigned markedly differing risks of dying after coronary artery bypass graft surgery, acute myocardial infarction, pneumonia, and stroke, depending on the severity measure used.[24] This resulted in conflicting impressions of relative hospital performance.

While such concerns are not always seen empirically—for example, other studies have found that hospital rankings remained stable irrespective of the severity measure used[25 26]—it is difficult to be confident about which severity adjustment scheme is the most valid. In any case, even when risk adjustment is carried out, the risk remains of confounding from significant unknown (and therefore unadjusted) factors. In a complex field like health care, the predictive power of even the best risk adjustment models will only ever be partial.

Risk adjustment is not only a complex problem in itself, but it is also a dynamic problem. The process of using risk adjusted figures to make comparisons over time is hampered by the problem of "upstaging"—that is, the grading of patients over time may shift upwards, perhaps because of greater attention being given to the initial assessment of severity. As the definitions of severity drift upwards, the highest risk patients in one category are moved up to the next highest risk category where they are lower risk relative to the other patients in the group. The highest risk patients from that group get moved up too, so each risk category loses some of its most severe cases and gains less severe cases. The outcomes (for example, mortality and morbidity) for each risk category considered separately thus appear to improve as the "pool" of severity within them is diluted. Such severity difference or "upstaging" has been seen in relation to coronary artery bypass graft surgery in New York State where some of the apparent improvements in mortality were in part attributed to drifts in severity assessments.[27] Guarding against upstaging in any given study may require periodic checking of the validity and reliability of any assessment tools to check for any drifts in grading practice.

Despite the considerable challenges involved in clinically credible risk adjustment, ignoring the issue is not an option: inadequate case mix adjustment can have a significant effect on the conclusions drawn from routine data. A review of seven published observational studies reporting a reduced mortality with increased volume of coronary artery bypass graft surgery found that the apparent benefit of receiving treatment in high volume hospitals decreased as the degree of case mix adjustment increased.[28] Furthermore, the size of the estimated benefit of treatment in a high volume centre reduced over time. Concluding that the estimates of benefit suggested in the literature are likely to be misleading because of inadequate adjustment for case mix, these authors and others[29 30] warn that other observational studies using routine data may overestimate the effect of high volumes of activity on the quality of health care.

Non-clinical factors can also have a major impact on quality measures, both in conjunction with case mix issues and independently of them. Yet adjusting for many of these factors can be harder than adjusting for clinical factors, leaving the possibility that apparently poorly performing hospitals may be penalised for factors outside their control. For example, in the US, community sociodemographic factors which impacted on patient physical and mental health status were found to have more influence on admission rates than physician practice styles.[31] Studies of admission rates to hospitals from different general practices in London show that much of the variation in hospital admission rates between GP practices is explicable by differences in patient populations, with a higher prevalence of chronic illness and deprivation being associated with higher admission rates.[32 33]

Added to these patient-specific characteristics which may affect quality assessments are confounding contextual factors over which health providers may have little or no control, such as the availability of services within that facility and in adjacent health or social care services. A study of the extent to which measures of population health, demographic characteristics, socioeconomic factors, and secondary care characteristics influenced admission rates from general practice[34] found that these factors explained substantial proportions of the variations between health authorities in admission rates for epilepsy, asthma, and diabetes (55%, 45%, and 33%, respectively). A further example exploring variations in cervical smear uptake rates among general practices found that there was marked variation between practices with rates ranging from 17% to 94%. Yet, using a multiple regression model, over half of this variation could be accounted for by patient and practice variables—notably the presence or absence of a female partner within the practice.[35]

In addition to adjusting for case mix, there is also therefore a need to think through broader system and contextual influences and how these may alter, diminish, or otherwise undermine the conclusions that can be drawn about variations. The outcome of such deliberations may lead to collection of a wider data set of potential confounders to allow exploration of some of these wider influences.

Coping with chance variability

Chance variability is present in all data and can hinder interpretation of routine data in two ways: (1) by showing apparent differences that are not real and (2) by obscuring real differences.

False alerts

Ranking or other comparisons arising from routine data contain de facto tests of hypotheses of difference. Yet statistical theory shows that when two units—for example, institutions or services—are compared, statistically significant differences will be seen one time in 20 (assuming the usual significance levels of 0.05 are used), even when there is no true difference between them. Routine data are therefore prone to identifying false outliers which may lead to hospitals being inappropriately praised or denigrated (the statistical "type I error"). This problem is obviously compounded when multiple comparisons are made, which is common in studies of healthcare quality. For example, in one study of acute myocardial infarction, Monte Carlo simulation modelling found that over 75% of hospitals assessed as "poor quality" on the strength of their high mortality rates for this condition were actually of average quality.[36] In a separate study evaluating the predictive power of early readmission rates in cardiac disease, around two thirds of hospitals labelled poor quality due to their outlier status on this measure were also found to be falsely labelled.[37] Even when providers are statistically worse than average, the extent of divergence not attributable to chance may be quite small. For example, death rates following congestive heart failure or acute myocardial infarction were 5–11% higher in one set of hospitals than in all other hospitals and yet, for each of these, most of the excess (56–82%) was compatible with purely random variation.[38]

The potentially misleading effects of multiple testing (leading to frequent type I errors) can largely be avoided by pre-specifying the key outcomes of interest and testing statistical significance only for these. Comparisons beyond this limited set are then treated more as hypothesis *raising* than hypothesis *testing*.

False reassurance

In addition to producing false outliers, chance can also hide real and important differences in quality within random fluctuations. Routine data are prone to providing false reassurance where important differences are not detected (statistically, the "type II error"). Thus, conventional statistical tests may fail to detect instances where hospital care is poor—for example, a hospital which fails to provide treatments which are known to be effective—because they concentrate on outcomes and these may appear within statistical norms when taken across disparate units. One study which used an analytical model to explore the sensitivity and predictive power of mortality rate indicators[39] found that fewer than 12% of poor quality hospitals emerged as high mortality rate outliers (indicating a high rate of type II errors), while over 60% of the "poor" outliers were in fact good quality hospitals (another high rate of type I errors).

For statistical reasons, differences in reported rates may or may not be statistically significant depending on how many years data are used.[23 34] Depending on what is being assessed, large patient numbers (and hence long time periods) may be required to be able to detect a difference. This may be inappropriate at the institutional level (as staff and facilities change) and impractical at the individual physician level. For example, in relation to a chronic disease like diabetes, it may be difficult to develop a reliable measure to assess physician performance and routine data may show little evidence either way.[40]

A priori power calculations can help determine how much data will be needed to say something meaningful about differences.[41] Such calculations made at the design stage can ensure that sufficient data can be collected to uncover potentially important differences. A crucial part of such a power calculation is the specification of what magnitude of difference would be regarded as *clinically* (as opposed to merely *statistically*) significant.

That outcome measures can be highly data demanding was well illustrated in a paper by Mant and Hicks[42] which used the management of acute myocardial infarction as an example to compare the relative sensitivity of measures of process and outcome in detecting deficiencies in care. The authors concluded that "*even with data aggregated over three years, with a perfect system of severity adjustment and identical case ascertainment and definition, disease specific mortality is an insensitive tool with which to compare quality of care among hospitals*".[42] The lesson that arises from such an observation is that systems designers should choose variables for which there is most likely to be a high degree of variability—which should lead to earlier emergence of any significant differences.

Real change or statistical artefact?

Comparisons of quality are not just about snapshots of practice; they are also intended to convey the extent of any changes in performance over time. Yet the problem of disentangling real change from statistical artefact is demonstrated in a study of the estimated adjusted live birth rate for

52 in vitro fertilisation clinics, the subsequent clinic ranking, and the associated uncertainty.[43] The researchers found that the confidence intervals were wide, particularly for those clinics placed in the middle ranks, and hence there was a high degree of uncertainty associated with the rankings. For example, one unit which fell in the middle of the rankings (27/52) had a 95% confidence interval for its rank ranging from 16 to 37. Furthermore, assessing changes in ranking in successive years also proved difficult. Looked at over two successive years, with only two exceptions, "*changes (in clinic ranking) of up to 23 places (out of the 52) were not associated with a significant improvement or decline in adjusted live birth rate*".[43]

Finally, the statistical phenomenon of "regression to the mean" will also account for some of the apparent movement of services up or down the rankings over a period of years[44]—so again not all observed changes are prima facie evidence of quality improvement or deterioration. While there are technical approaches that can, to some extent, adjust for such a phenomenon,[45 46] they are not simple and suggest the need for early statistical support at the design stage.

Data quality

The interpretation of comparative routine data thus faces three major challenges: (1) the need for appropriate measures; (2) the need to control for case mix and other variables, and (3) the need to minimise chance variability. These are all underpinned by a fourth challenge—namely, the need for high quality data. It is self-evident that poor quality data compound the inherent problems of interpretation of routine data described above and can undermine even the most sophisticated quality assessment tool. Yet, even in well resourced, well organised research studies, it is difficult to ensure that data are complete and of a consistently high quality.[9] It is harder still to ensure that routine data are of a high standard; the many factors that can compromise the reliability of both primary and secondary data have been well described.[23] For example, those collecting and entering the data are likely to be doing so as one of many daily tasks; data collection may be in conflict with other priorities and limited understanding of the purpose of the data may lead to unwitting errors in non-standard situations. Many different individuals will be involved in data collection and recording over long time periods and the risks of mistakes, inconsistencies or gaps are high. Thus, when routine data are used for quality assessment purposes, it is a common finding that they are inaccurate, with omissions and erroneous inclusions, or incomplete (especially in relation to some treatments) or insufficiently detailed for the purpose.[47]

Recent studies suggest that these difficulties persist. One study looked at whether the clinical conditions represented by the coding on hospital discharge summaries were actually present, as confirmed by clinical evidence in medical records. It found that, of 485 randomly sampled admissions, although there was clinical evidence to support most coded diagnoses of postoperative acute myocardial infarction, confirmatory evidence was lacking in at least 40% of other diagnoses, calling into question the interpretation of quality measures based on those diagnoses.[48]

Another recent study that looked at the accuracy of tumour registry data found that tumour registries provided largely accurate data on hospital based surgical treatment but there were large gaps in the data for outpatient treatments.[49] For example, the overall rate of radiation therapy after breast conserving surgery was found to be 80% when fresh data were gathered but was originally reported for only 48% of cases in the registry data. For adjuvant therapy the figures also diverged with less than one third of those who had in fact received adjuvant therapy having this information recorded within the registries. Tumour registries thus had significant omissions, particularly in failing to reflect adequately the care provided in outpatient settings.

Despite the importance of achieving accuracy in data collection, many studies have highlighted the fact that processes for routine data collection are still developing. One recent study which looked at 65 community health "report cards" in the US[50] found that there were significant variations across all areas, with data collection being the biggest challenge, and only half of the communities used pre-existing formats or the experience of others to guide report card development. The authors concluded that improved infrastructure and greater systematisation of the process would make it more sustainable. As these challenges appear to be significant in many settings, an early assessment of data quality is an essential first step for those seeking to exploit routine data sources for comparative studies of quality.

CONCLUDING REMARKS

Comparative data on the quality of health care serve many purposes and have the potential to both provide insight and drive quality improvement activities. Nonetheless, we have described a range of reasons why such judgements should be made with care. Deficiencies in measurement properties, problems with case mix, the clouding effects of chance, and the sometimes precarious nature of the underlying data sources all raise issues of concern (summarised in box 2).

Given the many problems surrounding the interpretation of routine data to assess health care quality, does this point towards the use of process measures or outcome measures? It is clear that concerns about data quality apply to both. Beyond this, it is the interpretation of outcome measures which is most susceptible to the serious threats posed by issues of validity and reliability, the confounding effects of case mix and other factors, and the problem of chance variability. For example, although process measures require the definition of appropriate processes of care for specific patient groups, the problem of case mix largely ends there. Adjusting for case mix in outcomes studies can be far more complex. Furthermore, whereas assessments of processes going awry can often be made from relatively small numbers, assessing such failures from outcomes alone requires much larger studies.[10 42] Process measures are also relatively easy to interpret, and they provide a direct link to the remedial action required. They may be particularly useful in revealing quality problems that are not susceptible to outcome measurement—for example, "near misses", unwanted outcomes, or unnecessary resource use.[10]

Another distinct problem with outcome measurement in terms of quality improvement and performance management is that, in many cases, the outcomes of interest are much delayed. While early postoperative mortality may be close to the point of intervention, many other key outcome measures are far removed from the period of clinical care. Re-operations for recurrent inguinal hernia or second joint replacement procedures, for example, usually occur late after surgery. Another pertinent example would be the use of survival to monitor the outcomes of cancer therapies, often measured several years after treatment. This is not to deny that these are important measures of outcome for patients and health systems. However, the fundamental problem in terms of performance management or quality improvement is that, by the time these measures are available, they will reflect clinical practice of several years previously and hence have somewhat limited capacity to influence. Thus, if quality assessment in health care is to mature, the enthusiasm for outcomes data will need to be tempered by due recognition of the complementary benefits of process data.

Clearly, many of the potential pitfalls highlighted in this paper can be overcome through (1) the development of indices with better metric properties, (2) more sophisticated risk adjustment, and (3) gathering more data and using alternative forms of statistical assessment such as control charts.[51–53] The success of each of these in turn rests upon high quality,

Box 2 Variations in measured quality: real difference or artefact?

When routine data reveal variations between different service providers in reported quality measures, this may be evidence of real differences in quality of care. Other possible causes of variation include:

(1) Problems with measurement. Validity and reliability of measures can be undermined by:

- inappropriate/insensitive data sources, e.g. data taken from administrative systems may lack necessary clinical details;
- measures which are too narrow to reflect the care provided, e.g. using a single measure in psychiatric care;
- inappropriate/insensitive definition of outcomes, e.g. looking at 30 day mortality for a particular condition even when most deaths fall outside this period;
- changes in data recording over time, e.g. apparent improvement or deterioration because of changes in reporting practices;
- differences in data recording between providers, e.g. data collection processes may not be equivalent and may lead to apparent variations;
- lack of blinding, e.g. unblinded assessment of outcome is prone to bias.

(2) The presence of case mix and other factors. Apparent differences between units may be more attributable to differences in the patient groups—for example, in clinical and sociodemographic terms and in terms of contextual factors—than to any true differences in performance. Yet case mix adjustment is demanding:

- it is always incomplete as adjustment can only be made when the necessary data are available and when the relevant factors are known;
- the choice of adjustment scheme can itself affect quality rankings;
- all adjustment schemes risk "upstaging" where the severity grading of patients may drift upwards over time, with implications for severity adjusted outcomes.

(3) Chance variability. This can lead to falsely identifying outliers for praise or blame (type I errors) or can obscure real differences and thus hide poor performers (type II errors).

(4) Poor data quality. Despite growing awareness of the problem, routine data systems are often incomplete or inaccurate, and this can seriously undermine conclusions drawn from such data.

locally derived and used, detailed clinical data sets.[47] Identifying the potential pitfalls, as we have done, highlights where most attention must be paid by quality researchers to ensure that the comparisons produced are robust and reliable indicators of real variations in quality practice.

Crucially, this paper has been concerned simply with the *interpretation* of apparent quality variations. We have not addressed the utility or utilisation of such comparative analyses, their proper role in local or national quality improvement efforts, or the scope for public involvement in their interpretation. Each of these issues is complex and contested.[54–57] Yet debates surrounding the use of such comparative analyses could be facilitated by a clearer understanding of the information content of routine data. It is to this end that the analysis presented is addressed.

Authors' affiliations

A E Powell, H T O Davies, Centre for Public Policy & Management, Department of Management, University of St Andrews, Fife, UK

Key messages

- Observational data assessing healthcare quality often show wide variations between geographical regions, health care providers, and even individual practitioners.
- Some of these variations may reflect real and important variations in actual healthcare quality, variations that merit further investigation and action.
- Apparent variation may also arise because of other misleading factors such as variations in sampling, differences in the validity and reliability of measures, or unadjusted case mix differences.
- Measures of process may be less susceptible to spurious variations than measures of outcome.
- Separating real from artefactual variations can be tricky, and judgements about the quality of care should therefore always be made with caution when using routine comparative data.

R G Thomson, Department of Epidemiology and Public Health, School of Health Sciences, Medical School, University of Newcastle, Newcastle upon Tyne, UK

Potential conflict of interest: Richard Thomson is the Director of the UK Quality Indicator Project, a branch of the International Quality Indicator Project, which provides anonymised feedback of comparative indicator data to UK hospitals.

REFERENCES

1 **Davies HT**, Marshall MN. Public disclosure of performance data: does the public get what the public wants? *Lancet* 1999;**353**:1639–40.

2 **Nutley SM**, Smith PC. League tables for performance improvement in health care. *J Health Serv Res Policy* 1998;**3**:50–7.

3 **Rainwater JA**, Romano PS, Antonius DM. The California Hospital Outcomes Project: how useful is California's report card for quality improvement? *Jt Comm J Qual Improve* 1998;**24**:31–9.

4 **Romano PS**, Rainwater JA, Antonius D. Grading the graders: how hospitals in California and New York perceive and interpret their report cards. *Med Care* 1999;**37**:295–305.

5 **Davies HT**. Public release of performance data and quality improvement: internal responses to external data by US health care providers. *Qual Health Care* 2001;**10**:104–10.

6 **Kohn LT**, Corrigan JM, Donaldson MS, eds. *To err is human: building a safer health system*. Washington, DC: National Academy Press, 2000.

7 **NCQA**. *HEDIS. The health plan employer data and information set.* Washington, DC: National Committee for Quality Assurance, 1999.

8 **Davies HT**, Crombie IK. Assessing the quality of care. *BMJ* 1995;**311**:766.

9 **Davies HT**, Crombie IK. Interpreting health outcomes. *J Eval Clin Pract* 1997;**3**:187–99.

10 **Crombie IK**, Davies HT. Beyond health outcomes: the advantages of measuring process. *J Eval Clin Pract* 1998;**4**:31–8.

11 **Campbell SM**, Braspenning J, Hutchinson A, *et al*. Research methods used in developing and applying quality indicators in primary care. *Qual Saf Health Care* 2002;**11**:358–64.

12 **Noseworthy J**, Ebers G, Vandervoort M, *et al*. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* 1994;**44**:16–20.

13 **Smith P**. On the unintended consequences of publishing performance data in the public sector. *Int J Public Admin* 1995;**18**:277–310.

14 **Leyland AH**, Boddy FA. League tables and acute myocardial infarction. *Lancet* 1998;**351**:555–8.

15 **Jenkins KJ**, Newburger JW, Lock JE, *et al*. In-hospital mortality for surgical repair of congenital heart defects: preliminary observations of variation by hospital caseload. *Pediatrics* 1995;**95**:323–30.

16 **Iezzoni LI**, Davis RB, Palmer RH, *et al*. Does the complications screening program flag cases with process of care problems? Using explicit criteria to judge processes. *Int J Qual Health Care* 1999;**11**:107–18.

17 **Stange KC**, Zyzanski SJ, Fedirko Smith T, *et al*. How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patient visits. *Med Care* 1998;**36**:851–67.

18 **Goldacre MJ**, Griffith M, Gill L, *et al*. In-hospital deaths as fraction of all deaths within 30 days of hospital admission for surgery: analysis of routine statistics. *BMJ* 2002;**324**:1069–70.

19 **Druss B**, Rosenheck R. Evaluation of the HEDIS measure of behavioural care quality. *Psychiatr Serv* 1997;**48**:71–5.

20 **Morgan K**, Prothero D, Frankel S. The rise in emergency admissions - crisis or artefact? Temporal analysis of health services data. *BMJ* 1999;**319**:158–9.

21 **Wakefield DS**, Hendryx MS, Uden-Holman T, *et al*. Comparing providers' performance: problems in making the "report card" analogy fit. *J Healthcare Qual* 1996;**18**:4–10.

22 **Hartz AJ**, Kuhn EM. Comparing hospitals that perform coronary artery bypass surgery: the effect of outcome measures and data sources. *Am J Publ Health* 1994;**84**:1609–14.

23 **Iezzoni LI**. Using risk-adjusted outcomes to assess clinical practice: an overview of issues pertaining to risk adjustment. *Ann Thorac Surg* 1994;**58**:1822–6.

24 **Iezzoni LI**. The risks of risk adjustment. *JAMA* 1997;**278**:1600–7.

25 **Shwartz M**, Iezzoni LI, Ash AS, *et al*. Do severity measures explain differences in length of hospital stay? The case of hip fracture. *Health Services Res* 1996;**31**:365–85.

26 **Iezzoni LI**, Shwartz M, Ash AS, *et al*. Does severity explain differences in hospital length of stay for pneumonia patients? *J Health Serv Res Policy* 1996;**1**:65–76.

27 **Green J**, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. *N Engl J Med* 1995;**332**:1229–32.

28 **Sowden AJ**, Deeks JJ, Sheldon TA. Volume and outcome in coronary artery bypass graft surgery: true association or artefact? *BMJ* 1995;**311**:151–5.

29 **Posnett J**. Is bigger better? Concentration in the provision of secondary care. *BMJ* 1999;**319**:1063–5.

30 **Spiegelhalter DJ**. Mortality and volume of cases in paediatric cardiac surgery: retrospective study based on routinely collected data. *BMJ* 2002;**324**:261–4.

31 **Komaromy M**, Lurie N, Osmond D, *et al*. Physician practice style and rates of hospitalization for chronic medical conditions. *Med Care* 1996;**34**:594–609.

32 **Reid FDA**, Cook DG, Majeed A. Explaining variation in hospital admission rates between general practices: cross sectional study. *BMJ* 1999;**319**:98–103.

33 **Majeed A**, Bardsley M, Morgan D, *et al*. Cross sectional study of primary care groups in London: association of measures of socioeconomic and health status with hospital admission rates. *BMJ* 2000;**321**:1057–60.

34 **Giuffrida A**, Gravelle H, Roland M. Measuring quality of care with routine data: avoiding confusion between performance indicators and health outcomes. *BMJ* 1999;**319**:94–8.

35 **Majeed F**, Cook D, Anderson H, *et al*. Using patient and general practice characteristics to explain variations in cervical smear uptake rates. *BMJ* 1994;**308**:1272–6.

36 **Hofer TP**, Hayward RA. Identifying poor-quality hospitals: can hospital mortality rates detect quality problems for medical diagnoses? *Med Care* 1996;**34**:737–53.

37 **Hofer TP**, Hayward RA. Can early re-admission rates accurately detect poor-quality hospitals? *Med Care* 1995;**33**:234–45.

38 **Park RE**, Brook RH, Kosecoff J, *et al*. Explaining variations in hospital death rates. Randomness, severity of illness, quality of care. *JAMA* 1990;**264**:484–90.

39 **Thomas JW**, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Med Care* 1999;**37**:83–92.

40 **Hofer TP**, Hayward RA, Greenfield S, *et al*. The unreliability of individual physician 'report cards' for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;**281**:2098–105.

41 **Florey CdV**. Sample size for beginners. *BMJ* 1993;**306**:1181–4.

42 **Mant J**, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *BMJ* 1995;**311**:793–6.

43 **Marshall EC**, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998;**316**:1701–5.

44 **Bland J**, Altman D. Statistics notes: some examples of regression towards the mean. *BMJ* 1994;**309**:780.

45 **Bland JM**, Altman DG. Some examples of regression towards the mean. *BMJ* 1994;**309**:780.

46 **Hayes RJ**. Methods for assessing whether change depends on initial value. *Stat Med* 1988;**7**:915–27.

47 **Black N**. High-quality clinical databases: breaking down barriers. *Lancet* 1999;**353**:1205–6.

48 **McCarthy EP**, Iezzoni LI, Davis RB, et al. Does clinical evidence support ICD-9-CM diagnosis coding of complications? *Med Care* 2000;**38**:868–87.

49 **Bickell NA**, Chassin MR. Determining the quality of breast cancer care: do tumor registries measure up? *Ann Intern Med* 2000;**132**:705–10.

50 **Fielding JE**, Sutherland CE, Halfron N. Community health report cards: results of a national survey. *Am J Prev Med* 1999;**17**:79–86.

51 **Mohammed MA**, Cheng KK, Rouse A, *et al*. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001;**357**:463.

52 **Adab P**, Rouse AM, Mohammed MA, *et al*. Performance league tables: the NHS deserves better. *BMJ* 2002;**324**:95–8.

53 **Lawrance RA**, Dorsch MF, Sapsford RJ, *et al*. Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice: observational study. *BMJ* 2001;**323**:324–7.

54 **Mannion R**, Davies HTO. Report cards in health care: learning from the past; prospects for the future. *J Eval Clin Pract* 2002;**8**:215–28.

55 **Davies HTO**, Bindman AB, Washington AE. Health care report cards: implications for the underserved and the organizations who provide for them. *J Health Politics Policy Law* 2002;**27**:379–99.

56 **Marshall MN**, Davies HTO. Public release of information on quality of care: how are the health service and the public expected to respond? *J Health Serv Res Policy* 2001;**6**:158–62.

57 **Lally J**, Thomson RG. Is indicator use for quality improvement and performance measurement compatible? In: Davies HTO, Tavakoli M, Malek M, *et al*, eds. *Managing quality: strategic issues in health care management*. Aldershot: Ashgate Publishing, 1999: 199–214.