The measurement of active errors: methodological issues

R J Lilford, M A Mohammed, D Braunholtz, T P Hofer

Qual Saf Health Care 2003;**12**(Suppl II):ii8–ii12

The value of research in any topic area turns on its validity. Patient safety research has revealed—or, at least, given renewed urgency to—a raft of methodological issues. The meaning and thus the value of empirical research in this field is contingent on getting the methodology right. The need for good methods for the measurement of error is necessary whenever an inference is intended and, since inferences lie at the heart of research and management, there is a huge need to understand better how to make measurements that are meaningful, precise, and accurate. In this paper we consider issues relating to the measurement of error and the need for more research.

A single patient safety incident may provide sufficient justification for management change. Thus, a single case of paraplegia from the inadvertent spinal injection of the wrong medicine or of surgery at the wrong site would provide the proper basis for management action without the need to measure incidence rates.

However, measurement of quality is often carried out for comparative purposes. There are two situations in which comparisons may be needed. The first is to compare organisations/clinicians, and the second is to draw cause and effect conclusions about how different policies or structures affect safety/quality. Both of these uses require inferences—the inference that states of affairs are different between one set of healthcare providers and another or that they have improved between one set of circumstances and another. Such inferences require that rates are measured and that they are measured both accurately and precisely. This paper is concerned with circumstances where such inferences may be called for and, hence, with measurement of rates. There are two ways that clinical quality can be measured/monitored: (1) measurement of outcome and (2) measurement of process.

MEASUREMENT OF OUTCOME VERSUS MEASUREMENT OF PROCESS

Outcomes can be broken down into mortality, physical morbidity, psychological well being, and satisfaction with services. Psychological outcomes, especially satisfaction, are likely to be sensitive to how care is delivered although, even in this context, comparisons may be misleading since expectations may vary systematically in place and time—some groups of patients may simply be easier to please. Monitoring is often used to measure physical outcomes—that is, mortality and morbidity. In some very particular circumstances such outcomes are a good reflection of the quality of care—for example, ureteric injury in laparoscopic gynaecological surgery. However, such cases are the exception and most mortality and morbidity is not the result of poor quality care. Most bad outcomes cannot be prevented or are side effects of treatment. If the factors that affect outcome and which cannot be controlled by an organisation/clinician vary systematically between organisations/clinicians, then comparisons are potentially biased. Outcome monitoring will always be needed to prompt research and non-judgemental investigation, but it should not form the basis for judgement or intervention except in situations where outcome really is a reliable measure of quality.

The important question is—to what extent are differences in outcomes a reflection of differences in the quality of care? It is widely accepted and agreed (and empirically demonstrated) that comparisons of outcomes must be adjusted for prognosis—for example, it would be inappropriate to compare hospital acquired infection rates between hospitals that carry out very different types of surgery. However, there are two ''risk adjustment fallacies''. The first is that risk adjustment might overadjust. Here, the quality related factor is associated with the risk factor so adjustment obscures real differences—for example, if age is a risk factor for death and older people are also cared for in a more perfunctory way, then adjusting for age might obscure real differences in treatment quality. The second fallacy, and arguably one that occurs far more frequently, is that risk adjustment may be insufficient. It is therefore incorrect automatically to attribute any differences in outcome after risk adjustment to quality of care; such differences may simply reflect differences in prognosis not captured in the dataset—either because the relevant prognostic variables were not measured or because they could not be measured. Thus, before we attribute differences in outcome to differences in quality of care, we need to know how much of the variance in the former is explained by variance in the latter. So what is the correlation between outcome and quality at the organisational level?[1]

The available evidence suggests that little if any of the differences in outcome between hospitals can be attributed to measurable differences in clinical process/active error rates.[1] This applies, for example, to many important clinical conditions such as treatment of heart attack,[2] pneumonia,[3] and cerebovascular accident.[2] This does not mean that the quality criteria/errors are not important or that they are not based on processes that have been shown to be effective. It simply means that the identifiable processes are one of many factors that affect outcome: the signal (outcome due to process) cannot be

See end of article for authors' affiliations

Correspondence to: R J Lilford, Department of Public Health and Epidemiology, University of Birmingham, Birmingham B15 2TT, UK; r.j.lilford@bham.ac.uk

distinguished from the noise (outcome due to other factors).[4] The argument that quality should be based on measurement of outcome because patients are interested in outcomes is specious. Patients want to maximise their personal outcomes and this cannot be achieved by misattributing cause and effect—that is, by misattributing group outcomes to particular institutions or policies. Not only does a system of punishment or reward based on outcome run a high risk of penalising or favouring the wrong providers, it also has little potential to improve health. Measurement of adverse events/outcomes can identify only those few institutions/clinicians who may lie outside some statistical threshold. However, demonstrating that a process is inadequate can lead directly to improvement irrespective of where an organisation lies in the ''league table'' of outcome—even those with better than average outcomes. Since most patients are treated in institutions whose outcomes lie in the normal range, a greater public health dividend lies in identification of error/quality (which can be used to shift the whole performance curve) than in spotting ''bad apples'' at the extreme of the outcome distribution.

There is also a scientific argument for concentrating on measurement of processes (of proven or undisputed importance). The sample sizes needed to show that management interventions are effective are far smaller if the outcome is a change in process than if it is an improvement in outcome.

Except in the case of the most egregious errors such as inadvertent injection of undiluted potassium chloride, most patients are unharmed by the errors to which they are exposed.[1 5 6] That is why large randomised trials are needed to define the correct standard of care in many instances; gains in relative risk of about 20% are typical of many healthcare interventions. Improving adherence to correct standards for heart attack from 60% to 80% might result in a difference in mortality of two percentage points (from 10% to 8%). The former could be demonstrated by examining 240 case notes (assuming power of 90% and $p<0.05$), while 1800 case notes would be needed to show the corresponding difference in mortality.

Thus, some error rates have a high risk of causing damage but they are very rare and therefore do not affect outcome rates at an institutional level. Others are common but have a much smaller individual impact[5] on outcome and are therefore likely to be lost in the ''noise''. It is therefore necessary to measure process—that is, clinical quality/active error rates[7]—for three reasons:

- in many, perhaps most, clinical situations outcomes are a poor reflection of quality of care and are therefore much more useful for guiding research and inquiry than for performance management involving judgement, reward and sanction;
- more gains in health care can be achieved by focusing on process (whereby everyone can improve by shifting the whole curve) than on outcome which leads to attention being focused on outliers;
- since error/protocol violations are much more common than the adverse events to which they predispose,[5] the statistical power tends to be orders of magnitude greater when processes, rather than outcomes, are compared.

MEASUREMENT OF THE PROCESS OF CARE

Hofer and colleagues argue that active errors and clinical process violations are one and the same.[8] By ''active errors'' we mean errors in patient care itself rather than in the system that may predispose to such errors. Thomas *et al*[9] have recently summarised methods for measuring error/process of care. In this paper we are concerned only with comparisons from which inferences may be made. This requires measurement of rates such as error rates. Reporting systems do not provide rates—they provide numerator information only—and we will not therefore consider them further. Processes of care may be measured by studying documentary (including computer) evidence or by observation.

Documentary data may be retrospective (examining data in computer systems or case notes) or prospective (by completing a pro forma as the case unfolds[10]). Observations may be made directly in real time or retrospectively by scrutiny of audio/video tapes. This paper is concerned mainly with documentary evidence, although many of the biases we describe apply to the various other methods. Whatever medium is used, quality of care may be assessed by two basic methods[11]—explicit and implicit.

By explicit we mean that the quality of care is assessed against predetermined criteria; an algorithm must be produced and the quality of care is then assessed against the criteria laid down. This method is sometimes referred to as ''criterion based assessment''. The implicit method is based on expert judgement and is not constrained by predetermined criteria. It is sometimes referred to as ''holistic judgement''.

Both of these methods have strengths and weaknesses (although they are not mutually exclusive). An obvious advantage of the explicit method is that it does not rely on expert judgement to the degree needed for implicit assessment. It also offers a method to protect against bias by expressing error rates in terms of the maximum number of errors possible in a data set (see below). The disadvantage of the holistic method is that it is poorly standardised (by definition) and expensive in terms of time and the skill level required. The advantage of implicit methods is that they can pick up processes of a diverse nature that would not be included in even the most complex of algorithms—for example, poor bladder catheter management in a patient with a heart attack.

MEASUREMENT ERROR IN THE MEASUREMENT OF ERROR

Medical record (case note/chart) review is the only method for which there are a substantial number of published estimates of reliability. Unfortunately, most of these are based on studies of adverse events and then judgement about whether the adverse event was caused by an error. The estimates are quite low (ranging from 0.2 to 0.3) for whether an adverse event was negligent or preventable (and up to 0.6 for whether an adverse event even occurred).[12] Goldman[13] reviewed 12 studies evaluating reliability (interobservational variation) of case note review and found a Kappa of only 0.4. Rubenstein[14] found a Kappa of 0.54 when two reviewers evaluated 333 charts, and Brennan[15] obtained a Kappa of 0.5 and 0.24 for adverse events and ''negligence'', respectively, when two reviewers were compared with a ''gold standard'' consisting of a team of ''super reviewers''. Hayward and Hofer[16] found an intraclass correlation of 0.19 between different reviewers of the same case notes.

Estimates of reliability are usually not calculated in a way which allows us to compare studies or to understand the relative contribution of reviewers, their training, or the difficulty of the decision task. We know that the more the heterogeneity in the raters and the conditions studied, the lower will be the reliability, but there is a need for more sophisticated studies of measurement properties that test a measurement procedure across a variety of different conditions of measurement. In particular, there is an urgent need to compare the explicit with the implicit methods and to focus on all errors, not just those associated with a poor outcome/adverse events.

Dean[17] reviewed the literature on drug administration errors and found that only one scale of severity, that of Bechtel *et al*,[18] had been assessed in the form of inter-rater reliability which was quite good (0.79). Delphi techniques can be used to agree error definitions in advance of studies of error rates,[19] and this is likely to be more valid than trying to reach group consensus after the data have been extracted, since this produces agreement within, but not between, groups.[20 21]

The presence of large amounts of measurement error can degrade the ability of researchers to measure the impact of interventions or provide evidence of association or causality between processes of care and outcome.

Quantifying measurement error and the sources of variation that cause it can, in some cases, allow investigators to account for it in the analysis and ultimately improve the measurement process. We hypothesise that explicit measurement of predefined error will be much more reliable than implicit assessment, but that it will miss more errors. Moreover, we hypothesise that previous findings that there is little correlation between errors of different types within an organisation will be confirmed.[5] The NHS Research Methods Programme has issued a call for a proposal to test these hypotheses. There is little information about the measurement properties of any of the other methods (see below) for measuring error, many of which have substantial theoretical advantages over physician chart review in detecting error.

BIAS IN MEASUREMENT

Directly measuring errors, rather than inferring from adverse events, is not a panacea since errors are subject to case mix considerations (as are adverse events) because different patients may have different opportunities for error. There are two approaches to this problem. Firstly, a possibility not easily applicable to adverse events is to express errors in terms of direct enumeration of ''opportunity for error'', rather than using ''patient'' as the denominator. However, this is possible only when using a predetermined pro forma algorithm that defines such ''opportunities''. Secondly, per patient error rates can be used with an attempt at statistical adjustment, but it must be recognised that there is no infallible risk adjustment method.

There is also a risk of what we may term ''information bias''. By this we mean that the diligence with which information is recorded may influence the ''visibility'' of errors. Clinicians of high diligence may record more data and thereby expose themselves to detection of more ''violations''. To guard against this we recommend that errors are classified as primary and secondary—primary relate to record of observations that should be present and secondary to contingent actions. This does not eliminate, but should ameliorate, the problem. Consider, for example, the management of community acquired pneumonia. Recording oxygen saturation would be a primary process, while responding to falling oxygen saturation levels would be a secondary process. A particular type of information bias arises when an intervention designed to reduce error interacts[5] with the measurement method. For example, computer systems designed to improve care may affect the recording of information in case notes and hence the proportion of errors that are detected by case note review. There is an urgent need, therefore, to compare different methods of assessing quality/error rates when computer systems are introduced.

Observer bias is another theoretical risk. It is extremely expensive to mask notes to enable blind measurements on case note review (Jocelyn Cornwall, acting chief executive CHI, personal communication). One method that has been used is to re-dictate the notes, but important information may be left out. Perhaps the most practical method is to try and ensure that observers are blind to the hypothesis being tested.

Two further points are worth making. Firstly, when studying the effectiveness of methods to improve care, differences in performance on a particular criterion are often large between organisations.[5] Moderate biases are much less likely to account for large observed differences in process than the typically smaller differences in outcome. Secondly, with respect to performance management, some bias in process management is arguably less important than in the case of outcome because it can be corrected directly. It is clear where the problem lies so that it can be tackled head on and the organisation/clinician need not be stigmatised by a finding whose cause is opaque and remedy uncertain.

SENSITIVITY AND SPECIFICITY OF DIFFERENT METHODS

Further empirical work is needed to compare error detection rates and statistical properties of different methods. Stanhope and colleagues[22] showed that many adverse events are neither reported nor captured on case notes. For example, Michel *et al*[10] used the retrospective method to compare prospective and cross sectional methods based on data collection from clinical units for assessing preventable adverse event rates in acute hospitals. The results suggested that the prospective method is more sensitive—that is, it detects a greater range and number of events.[10] However, prospective data collection by clinicians may be subject to bias. In before and after studies the clinicians are both the subject of change and the observers of the effect of that change. In comparative studies better clinicians may be more sensitive in spotting errors/adverse events and hence may make themselves look worse. Perhaps the gold standard is unobtrusive direct observations made with appropriate consent[23] by third party observers blind to the hypothesis being tested. However, this method is extremely expensive, especially in low error environments.

DEVELOPING NEW METHODS TO MEASURE ACTIVE ERROR

It is possible (perhaps likely) that only a minority of clinical errors are captured by existing methods. Thus, while certain clear cut violations such as failure to check a patient's blood potassium level when indicated or failure to recognise well established signs of meningitis can be detected from, for example, case note review, other factors such as the quality of communication with patients or surgical skill remain largely in the tacit domain—that is, without adequate measures or even definitions. The metric properties of these measures (many of which will need to be multidimensional) will need to be assessed—that is, their validity and properties established. For example, surgical skill could be assessed by theatre sisters or assistants and blind peer review of videotaped operations and the correlations between these methods measured. If these were then found to be correlated, further studies would be possible to find out whether methods which improve surgical skill correlate with outcome—for example, wound infection, recovery time—and hence test criterion validity. It would also then be possible to measure how training in simulations transfers into the clinical environment, how aptitude correlates with time in training and skill level achieved, etc.

Some have suggested creating a clinical practice ''black box'' with inbuilt algorithms based on ''triggers'' to identify and direct attention to a sequence in the care pathway where errors might have occurred. In addition to methods based on videotaping and observing patient care and the systems that underpin such care, further ideas need evaluation. These include ''participant ethnography'' in the tradition of

Pointers for future research

- A large scale comparison of explicit and implicit case note review is urgently needed; the NHS Research Methods Programme has issued a call for such a study.
- This should be repeated in many different settings.
- A comparative study of different methods to measure error rates should be undertaken, but it must be borne in mind that the method which identifies most errors is not necessarily the least biased.
- Experiments are needed to develop new methods to measure active errors, particularly of practical procedures such as surgery and resuscitation.
- Many new methods for error assessment such as simulated patients require more detailed assessment.
- Once better measurements of active errors have been produced, the effect of latent factors (systems factors such as culture, morale, etc) on error rate and on outcome should be measured.

Rosenham[24] which could be developed and evaluated.[25] Such methods may include the use of "standardised patients"[26] (actors playing out the role and patients themselves). Again, measurements of reliability and comparisons with other methods are underdeveloped, although examples exist.[27]

Since standardised patients cannot be used for many aspects of care (such as undergoing surgery or intensive care), other methods also need to be developed and evaluated. In this context it should be possible to develop methods to elicit the experiences and observations and narratives of patients that could help in identifying threats to patient safety. The use of individuals with considerable familiarity with the particular disease (either as sufferers or carers) or clinicians who become victims of the disease in which they specialise is a topic requiring further research. It will be important not only to confirm interobserver variation within these methods, but also to see how methods correlate—for example, well informed standardised patients *v* case note review *v* prospective audit.

Another topic for further research is how appropriately to combine results from different measurement methods into one composite measure. For example, patient safety studies could simultaneously use direct observations, chart review, and incident reports to develop one overall measure of error/adverse events. What is the correct way to combine these measures? Since research will involve comparison of different methods to study rates of error, it is likely that methods of triangulation developed in the natural and social sciences[28] will need to be adapted and evaluated for the specific context of patient safety.

CONCLUSION

The study of measurement of error is in its infancy, perhaps because people have (wrongly) thought that measurement of outcome/adverse events/harm would suffice. Developing and validating measurement methods is a formidable undertaking but, if the quality movement/error science is to reach its potential, this should be made a priority.

Authors' affiliations
R J Lilford, M A Mohammed, D Braunholtz, Department of Public Health and Epidemiology, University of Birmingham, Birmingham B15 2TT, UK
T P Hofer, HSR&D (11H), VA Ann Arbor Healthcare System, 2215 Ann Arbor, MI 48105, USA

Key messages

- Case note (chart) review is by far the most thoroughly studied method for measurement of errors.
- Interobserver agreement is reasonable for detection of an adverse event but poor for deciding whether it was caused by an error.
- Explicit methods of error detection are likely to have much better interobserver agreement but also considerably less sensitivity than implicit methods.
- Explicit methods can be based on opportunity for error, not patients, and hence overcome, or at least ameliorate, error due to differences in case mix.
- Prospective data collation by clinical staff was shown to be surprisingly sensitive in the single example where it was compared with case note review, but it is subject to observer bias if institutions or policies are compared.
- Prospective data collation by third party observers has revealed errors not detectable by case note review but it is expensive and difficult to arrange ethically in some settings.
- Outcomes should not be used to penalise or promote clinicians or institutions.
- Process measurement enables a whole service to improve because most providers lie in the middle of any distribution and more gain can be achieved by shifting the mean than by truncating the tail.

REFERENCES

1 **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.
2 **Park RE**, *et al*. Explaining variations in hospital death rates. Randomness, severity of illness, quality of care. *JAMA* 1990;**264**:484–90.
3 **Dubois R**, Rogers WH, Moxley J, *et al*. Hospital inpatient mortality: is it a product of quality? *N Engl J Med* 1987;**317**:674–80.
4 **Ash A**. Identifying poor-quality hospitals with mortality rates. Often there's more noise than signal. *Med Care* 1996;**34**:735–6.
5 **Wilson B et al**. The Leeds University Maternity Audit Project. *Int J Qual Health Care* 2002;**14**:175–81.
6 **Mant J**. Process versus outcome indicators in the assessment of quality of health care. *Int J Qual Health Care* 2001;**13**:475–80.
7 **Lilford RJ**, Mohammed MA, Spiegelhalter D, *et al*. Performance monitoring of health care organizations-to judge or not to judge the quality of care? *Lancet* 2003; (submitted for publication).
8 **Hofer TP**, Kerr EA, Hayward RA. What is an error? *Effect Clin Pract* 2000;**3**:261–9.
9 **Thomas EJ**, Petersen LA. Measuring errors and adverse events in health care. *J Gen Intern Med* 2003;**18**.
10 **Michel P**, Quenon JL, Sarasqueta AM, *et al*. L'estimation du risqué iatrogene graves dans les establissements de santé en France: les enseignements d'une étude pilote dans la region Aquitaine. *Etudes et Resultats* 2003;**219**:1–8.
11 **Brook RH**, Appel FA. Quality-of-care assessment: choosing a method for peer review. *N Engl J Med* 1973;**288**:1323–9.
12 **Thomas EJ**, Lipsitz SR, Studdert DM, *et al*. The reliability of medical record review for estimating adverse event rates. *Ann Intern Med* 2002;**136**:812–6.
13 **Goldman RL**. The reliability of peer assessments of quality of care. *JAMA* 1992;**267**:958–60.
14 **Rubenstein LV**, Kahn KL, Reinisch EJ, *et al*. Changes in quality of care for five diseases measured by implicit review, 1981 to 1986. *JAMA* 1990;**264**:1974–9.
15 **Brennan TA**, Localio RJ, Laird NJ. Reliability and validity of judgements concerning adverse events suffered by hospitalised patients. *Med Care* 1989;**27**:1148–58.
16 **Hayward RA**, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA* 2001;**286**:415–20.
17 **Dean B**. *Hospital medication administration errors: their simulation, observation and severity assessment*, Thesis/dissertation, 1999:3–314.
18 **Bechtel GA**, Vertrees JL, Swartzberg B. A continuous quality improvement approach to medication administration. *J Nurs Care Qual* 1993;**7**:28–34.
19 **Dean B**, Barber N, Schachter M. What is a prescribing error? *Qual Health Care* 2000;**9**:232–7.
20 **Gigone D**, Hastie R. The common knowledge effect: information sharing and group judgement. *J Personality Soc Psychol* 1993;**65**:959–74.

21 **Hofer TP**, Bernstein SJ, DeMonner S, *et al*. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Med Care* 2000;**38**:152–61.
22 **Stanhope N**, Crowley-Murphy M, Vincent C, *et al*. An evaluation of adverse incident reporting. *J Eval Clin Pract* 1999;**5**:5–12.
23 **Donchin Y**, Gopher D, Olin M, *et al*. A look into the nature and causes of human errors in the intensive care unit. *Crit Care Med* 1995;**23**:294–300.
24 **Rosenham DL**. Being sane in insane places. *Science* 1975;**179**:250–8.
25 **Verho H**, Arnetz JE. Validation and application of an instrument for measuring patient relatives' perception of quality of geratric care. *Int J Qual Health Care* 2003;**15**:197–206.
26 **Beaulieu N**, Rivard M, Hudon E, *et al*. Using standardized patients to measure performance of physicians. *Int J Qual Health Care* 2003;**15**:251–9.
27 **Peabody JW**, Luck J, Glassman P, *et al*. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA* 2000;**283**:1715–22.
28 **Bryman A**. *Social research methods*. Oxford: Oxford University Press, 2001.