Current perspectives in assessment: the assessment of performance at work

John J Norcini

BACKGROUND Traditional assessment has improved significantly over the past 50 years. A number of new testing methods are now in place, the computer is improving both the fidelity and efficiency of examinations, and the psychometric principles on which assessment rests are more sophisticated than ever.

AIM There is growing interest in quality improvement and there are increasing demands for public accountability. This has shifted the focus of testing from education to work. The purpose of this paper is to describe the assessment of work.

DISCUSSION In contrast to traditional assessment, there are no 'methods' for the evaluation of work because the content and difficulty of the examination are not controlled in any fashion. Instead it is a matter of identifying the basis for the judgements (outcomes, process, or volume), deciding how the data will be gathered (practice records, administrative databases, diaries/logs, or observation), and avoiding threats to validity and reliability (patient mix, patient complexity, attribution, and numbers of patients).

FUTURE DIRECTIONS Overall, the assessment of doctors' performance at work is in its infancy and much research and development is needed. Nonetheless, it is being used increasingly in programmes of continuous quality improvement and accountability. It is critical that refinements occur quickly to ensure that patients receive the highest quality of care and that doctors are treated fairly and provided with the information they need to guide their professional development.

KEYWORDS physician family/*standards; clinical competence/*standards; quality control; quality of health care/*standards

*Medical Education 2005; **39**: 880–889*
doi:10.1111/j.1365-2929.2005.02182.x

INTRODUCTION

For the first half of the 20th century, assessment in medical education consisted largely of written and oral examinations, with medical students and trainees generally being the people tested. The written examinations were mostly essays and the oral examination often involved the long case.

From the 1950s to the present, there have been significant changes on 3 fronts. First, a plethora of new methods of assessment has been developed, with the goal of measuring all aspects of doctor competence. For instance, several different types of written questions, including multiple-choice questions (MCQs), are aimed at testing knowledge, the objective structured clinical examination (OSCE) and its variants provide a means of assessing clinical skills, and there are now a variety of techniques for addressing non-cognitive competencies.[1–3]

Secondly, the computer has become an integral part of testing. At first, it was used to scan and score large-scale MCQ examinations. Over time, however, its role has become more central. On the one hand, the computer's intelligence has been put to use in selecting which questions will be administered to particular students. Adaptive testing based on item response theory permits gains in efficiency and precision.[4,5] On the other hand, computers have become tools for high

Foundation for Advancement of International Medical Education and Research (FAIMER), Philadelphia, Pennsylvania, USA

Correspondence: John J Norcini PhD, Foundation for Advancement of International Medical Education and Research (FAIMER®), 3624 Market Street, 4th Floor, Philadelphia, Pennsylvania 19104, USA. Tel: 00 1 215 823 2170; Fax: 00 1 215 386 3309; E-mail: jnorcini@faimer.org

Overview

What is already known on this subject

Traditional assessment has improved significantly over the past 50 years.

The growing interest in quality improvement bolstered by increasing demands for public accountability has shifted the focus to an assessment of work.

What this study adds

The assessment of work requires identification of the basis for judgement, decisions on how the data will be gathered, and avoidance of threats to validity and reliability.

Suggestions for further research

The assessment of work performance is in its infancy and much research is needed.

Among the threats to validity and reliability, the development of good risk adjustment procedures is a priority.

Determining the number of patients necessary to achieve good estimates in support of a doctor's assessment is also needed.

fidelity recreation of aspects of the clinical encounter. For instance, there are now very good simulators for a wide range of procedures and patients.[6,7]

Thirdly, research on assessment in medical education has identified several sources of error in the examinations given and has proposed ways to reduce that error. For example, doctor performance was found to be case- or patient-specific, so in order to develop a stable estimate of an examinee's performance, several cases or patients need to be sampled.[8] In addition, methods have been developed to handle routine measurement challenges such as equating and standard setting.[9,10] Finally, the psychometric theory underlying assessment has been further developed.[11,12]

Throughout this period, the primary object of assessment has been the student or trainee and the goals of assessment have typically been to support the educational process and/or to establish the competence of individual doctors. There certainly remain unresolved issues and challenges in this form of assessment, both in terms of the individual methods and in creating an integrated system for them.[13,14] However, significant progress has been made.

Starting in the early 1990s, efforts to improve the quality of health care increased in intensity. These were driven by rising costs, concerns over patient safety, and public demands for accountability. The efforts have relied on methods developed by workers in the field of quality management science and used successfully in industry for a number of years.[15,16]

The assessment of actual performance in practice is essential to the quality management sciences. The principal measures for doctors are patient outcomes and the process of care that they provide as part of their routine work. These serve to identify areas for improvement, signal whether goals have been achieved, and respond to the public's increasing demands for accountability.[17]

Because of the events of the past decade, assessment faces a new set of challenges. The venue has moved from the relatively controlled and homogeneous settings of education to the uncontrolled and heterogeneous world of work. Instead of students and trainees with identified supervisors/teachers and the expectation of assessment, the object of measurement has shifted to practising doctors who are supervised to a far lesser degree and who find the prospect of assessment considerably less appealing. This creates significant new challenges for the conduct and meaning of assessment.

THE ASSESSMENT OF PERFORMANCE AT WORK

In traditional assessment, the developers select a method with known characteristics, create test material for it, and then directly control and manipulate all aspects of its application. This ensures that:

1. the content maps onto the domain to which the results should generalise;
2. the test is long enough to produce reliable scores;
3. performance is wholly attributable to the examinee, and
4. different versions of the test are comparable in difficulty, enabling comparisons among examinees and against standards.

In contrast to traditional testing methods, which control the stimuli to which examinees respond, an assessment of the work of doctors must be based on their responses to the patients they see. The patients are analogous to the stems of MCQs or the stations of an OSCE, and the content of the test is whatever challenges those patients present. Rather than creating and using the best methods of assessment, as is traditional in educational measurement, the task in the assessment of work is to take whatever patient data exist and then adjust, eliminate and aggregate them to produce scores that have meaning in the comparison of doctors between doctors and/or against external standards. Consequently, there are no 'methods' for assessing work in the traditional sense. Instead, there are decisions about the basis for the assessment of the doctors, the sources of information that support those assessments, and the threats to the validity and reliability of the results.[18–20]

Basis of the assessment

The principal measures of performance in health care systems are patient outcomes, the process of care that doctors provide, and the volume of services they offer. These can also serve as the basis for assessing individual doctors.

Patient outcomes

The assessment of individual doctors can be based on judgements about the outcomes of their patients. Historically, these outcomes have been mortality and morbidity. For instance, the quality of care provided by orthopaedic surgeons can be judged in part by the mortality of their patients following hip replacement. Likewise, the quality of care provided by endocrinologists can be judged in part by the rate of lower extremity amputation in patients with diabetes.

Many view patient outcomes as the best measure of health care quality. They serve as the ultimate measure of accountability to the public and provide reassurance that a doctor is performing well.[21] For doctors themselves, it offers the fairest measure of their competence as the assessment is based on their own actions within the context of their practice. This stands in contrast to traditional forms of assessment, which can be tailored only crudely to what doctors actually do. For health care systems, outcomes can identify effective doctors who can improve both the quality and value of care. Finally, for patients, publication of the data serves as a basis for choosing among providers when options are available and it often leads to improvement in quality.[22]

In recent years, the number of outcomes used to judge the quality of care has expanded significantly. In addition to mortality and morbidity, physiological measures (e.g. blood pressure), clinical events (e.g. stroke), symptoms (e.g. difficulty in breathing), functional status, patient satisfaction and experiences with care, and cost effectiveness are now being used as the basis for judging the outcome of doctors, hospitals and treatments.

Process of care

Another way of assessing doctors is based on judgements about the process of care that they provide to their patients. Screening is one such basis for making these types of judgements. For example, the US Preventive Services Task Force recommends that clinicians routinely screen men aged 35 years and older and women aged 45 years and older for lipid disorders.[23] Doctors are judged according to how many of their patients who meet these qualifications have been screened, with the expectation of 100%.

In addition to screening, there are a number of other general process measures such as preventive services, immunisations, patient education and counselling. For example, there is a series and schedule of recommended immunisations for infants and children.[24] Similarly, there are preventive services that should be delivered routinely in the primary care setting.[25] For all of these, doctors can be judged on the percentage of their patients who receive the services.

More recently, disease-specific process measures have received attention. For diabetes patients, doctors may be judged on whether they have routinely measured HbA_{1c} and completed foot and eye examinations. For pneumonia, they may be assessed on the percentage of patients who had blood cultures before antibiotics were administered and who received influenza screening or vaccination.

Assessment based on process measures has a number of advantages, chief among them being that they are directly within the control of the doctor and fit well with continuous quality improvement programmes. The major disadvantage is that doing the right thing does not guarantee the best outcomes for patients. For instance, the fact that doctors screen their patients for lipids does not mean that they make the right decisions regarding treatment for those who require it.

Volume

A final way of assessing doctors is based on how often they provide particular services to their patients. For example, an assessment of surgeons might involve checking how often they perform total hip replacements. Likewise, an assessment of cardiologists might include gathering information on the number of patients with acute myocardial infarction they have treated within the past year.

The basis for assuming that simple counts provide for valid assessment is the relatively large literature indicating a relationship between provider volume and quality of care.[26] For example, patients of low-volume surgeons have higher rates of revision of total hip replacement than do patients of high-volume surgeons.[27] Similarly, cardiologists who treated a greater number of patients with acute myocardial infarctions had lower 30-day mortality rates than cardiologists who treated smaller numbers.[28]

The advantage of using volume as an indicator of quality of care is that the data are easy to obtain and comparisons among doctors are meaningful and straightforward. The disadvantage is that the fact that something is being done does not necessarily provide reassurance that it is being done right.

Sources of information for the assessment

Assessment information can be obtained from at least 4 sources: clinical practice records, administrative databases, diaries or clinical logs, and observation. Each has different strengths and weaknesses in terms of supporting judgements based on outcomes, process and volume.

Clinical practice records

The clinical practice record is an account of the patient's history. It contains findings, test results and treatment information. As such, it ranks among the best sources of information on patient outcomes, the process of care, and volume. To derive data from these records, information is usually extracted by one or more individuals who are trained to do so.

Unfortunately, the audit process is extremely expensive and time-consuming, especially when the records of a sizeable number of patients are involved. It is also made less than ideal by the fact that records are often incomplete and illegible. However, training, monitoring and feedback have been successful in improving the quality of the data obtained from practice records.[29,30]

Although it remains several years away, the electronic medical record is the ultimate solution to this problem. It offers rapid access to the data and it allows manipulation of the results with relative ease. In the meantime, self-audit of clinical practice records is a credible alternative as long as it is coupled with an external audit of a sample of the doctors being assessed.

Administrative databases

Considerable data about the practice of doctors is generated and stored as part of the process of administering health care and reimbursing for services. The clinical content of such databases is typically limited to demographics, diagnoses, and codes for procedures. Such data are easily available, inexpensive, and based on large populations of patients and doctors. They have been used successfully to support research and quality improvement efforts.[31]

While not nearly as rich as clinical practice records, administrative databases can be used as a source of information on patient outcomes, process and volume for individual doctors. However, the sizeable gaps in clinical information do not permit a full understanding of whether the care was appropriate and whether there were errors in judgement. Moreover, data collected for purposes of billing may not accurately reflect the nature and range of services provided. Hence, administrative data may best be used as part of a screening process in the assessment of doctors.

Diaries or case logs

Doctors, especially those in training, sometimes keep a listing of the procedures they perform or the patients they see. Procedure logs often contain information on which procedure was performed, who observed it, whether there were complications, information about the patient, and data on whether the procedure was performed adequately. Case or patient logs might gather data concerning the patients' demographics, their referral sources, their main problems, what the trainee did, the observer, and an evaluation of the performance. Increasingly, these logs are moving from paper to web-based or hand-held device-based systems to enhance their efficiency.[32,33]

This is a reasonable way to collect volume data, but it is less useful for process and patient

outcomes. In many settings, the doctors themselves choose which patient or procedure becomes part of their diary and who observes the performance. This self-selection has adverse consequences for the validity of the assessment as important content might be omitted and the observers may be chosen to provide more favourable evaluations.

Observation

There are many ways of collecting data through observations but they need to be routine or covert to qualify as a reflection of the quality of work. Otherwise, if the doctors know they are being assessed, it may alter their performance. In turn, this would make the results less likely to generalise to daily activities.[34]

The most common observers are supervisors, peers, patients, and other health care providers, while ratings are the most typical way of gathering this type of data. When collecting it, there are at least 4 issues that should be considered. Firstly, the expertise of the observer should match the judgements they are being asked to make, assuming their task is not simply to note the occurrence of particular behaviours. For example, doctors should collect data about and evaluate other doctors on medical matters. Similarly, patients are a better source of information and evaluation about communication skills. Secondly, it is useful to have different observers contribute to the assessment of a single doctor. This enhances the validity of the data and also increases its reliability. Thirdly, training of the observers is necessary depending on the complexity of the task, but even extensive training is not a substitute for adding observers. Fourthly, the relationship between the observer and the doctor may adversely influence the validity of the assessment.[35,36]

Observation is best suited to collecting data concerning the process of care. It is less useful for gathering information on patient outcomes or volume.

Threats to the validity and reliability of the assessment of performance at work

There are a variety of threats to the validity and reliability of assessments of performance at work, including patient mix, patient complexity, attribution and numbers of patients. Each is described below and analogies are drawn to traditional assessment methods.

Patient mix

Doctors have a unique panel of patients who differ in the mix of problems they present.[37,38] This variability comes from a host of sources, including the doctor's specialty, the nature and location of the practice, and self-selection. Even for doctors in the same specialty, there is considerable variation in the nature of the patient problems they encounter. Some doctors treat patients for whom the prognosis is quite poor, while others treat patients where the clinical outcomes are likely to be good. In addition, patients with specific characteristics (e.g. age or health status) choose doctors with particular characteristics (e.g. practice style, proximity).

In a typical educational assessment, this would be analogous to mounting an OSCE composed of varying numbers stations, all of which were unique to the student, drawn from different content areas, and of unequal difficulty. Furthermore, some proportion of the stations would be self-selected and some would be assigned in a non-random fashion. The fact that every student's OSCE would be composed of a different mix of cases poses 3 problems. Firstly, the results of the assessment will only generalise to performance with a panel of patients having the same set of conditions. This means that the results will be predictive of future performance for a limited, unique domain. Secondly, it is difficult to compare 1 student to another, as no 2 students have the same patients with the same problems. Thirdly, the standards of performance will vary with the content, so standards must be developed for every version of the assessment as each is unique.

Similarly, in an assessment of work, differences in patient mix are completely confounded by the doctor, making it difficult to disentangle his or her unique contribution to patient outcomes and the process of care provided.[39] Therefore, comparisons against each other or to a set of standards are prone to bias. Without a sampling strategy, these biases are so large as to make the results unsuitable for use in a quality assessment or improvement programme.[40,41]

The only practical way to address these problems is to focus the assessment on a particular condition (sometimes called a tracer condition) and include only those patients who have that problem.[42] For example, considerable work has been carried out on diabetes.[43] The American Diabetes Association recruited groups of experts who identified both process of care measures (e.g. lipid profiles, eye examination) and patient outcomes (e.g. HbA_{1c},

satisfaction) that formed the basis for making judgements about doctors. These measures can be aggregated to produce a score that reflects the quality of care an individual diabetes patient has received and then aggregated again at the level of the doctor to determine his or her overall performance.

It is important that the condition be common so that the same assessment can be conducted for a sizeable number of doctors. In addition, the type of condition should be chosen such that the doctor can make a difference and so that the consequences of intervention are significant in terms of patient outcomes.

Focusing the assessment on individual conditions overcomes some of the biases associated with patient mix and ensures, from the perspective of accountability, that the doctor is able to produce reasonable outcomes for the most frequent and important medical conditions. However, it does not provide a comprehensive picture of a doctor's competence and it does not permit the assessment of conditions that are important but relatively infrequent. It is also unable to address patient problems where scientific and clinical advances are changing the nature of practice.

Patient complexity

Even patients with the same condition will vary widely depending on a host of factors. The process of care rendered by the doctor and the patient's outcomes are obviously influenced by the severity of the illness.[44] These are also affected to a considerable degree by the patient's other problems (comorbidities).[45] Both severity of illness and comorbidities are often captured in the medical record and can be taken into account when making judgements about individual doctors.

Unfortunately, a number of undocumented factors also affect patient outcomes. For instance, patient adherence to treatment plans can vary widely depending on financial resources, willingness and ability to comply. This problem is exacerbated by the fact that some doctors tend to take on the most challenging patients, while others refer them to more specialised colleagues or systems of care.

In a typical educational assessment setting, this is analogous to giving each student in the class a set of unique MCQs on the same topic. Some of the MCQs would be self-selected and others would be assigned in a non-random fashion. This distribution of items makes unclear the meaning of the scores on the test. A high score may well reflect the fact that a student knows more than his classmates with low scores. However, he may have been better at selecting items that were well suited to his strengths or may have been assigned an easier set of items.

Similarly, in a work setting, patient complexity is completely confounded with the doctors, making it difficult to know the degree to which they were responsible for the patients' outcomes. Hence, comparisons among doctors or against a set of standards are likely to be biased. Without some form of adjustment, these biases are so large that it is inadvisable to use patient outcomes in the setting of a quality assessment or improvement programme.

One way of addressing the problem of patient complexity is by excluding those patients who are gravely ill because of a comorbid condition. For example, in the case of acute myocardial infarction, it may be reasonable to eliminate patients with anoxic brain damage, metastatic cancer or extensive trauma. A related option is to eliminate patients who have 2 or more comorbid conditions. This will restrict the range of patient complexity although substantial differences may remain, but it will also reduce the number of patients available to make judgements about each doctor.

A more refined approach to the problem of patient complexity is to apply a risk adjustment to patient outcomes based on comorbidities, severity of illness, or both.[46–48] The first step is to collect clinical data on a sizeable group of patients with the same condition. Included would be demographic, comorbid, symptom and laboratory data as well as outcomes such as mortality or quality of life. These are combined statistically to produce the best prediction of patients' outcomes given their clinical status. Doctors are not judged by how well their patients do in an absolute sense, but how well they do against these expectations.

For example, to develop a risk adjustment for 30-day mortality following myocardial infarction an investigator would collect information for a large group of patients who had this diagnosis. Included would be demographic information like age, sex, and socioeconomic status, comorbid conditions such as diabetes, hypertension and renal failure, and medical aspects of the event itself such as the site of the infarct, results of the electrocardiogram, or cardiac enzymes. These data are then combined using one of a variety of statistical methods to produce a predicted probability of death. Commercial risk adjustments are available for many of the most frequent and important medical conditions.[49]

Although some form of risk adjustment is essential, there are problems with the procedures currently being used. It is difficult to identify all of the factors that influence patients' outcomes, the statistical corrections themselves are relatively crude, and there is variation due to the sample of patients, doctors and institutions used to develop the model. Consequently, different risk adjustments disagree about the probability of mortality or morbidity for the same patients.[50] Likewise, they disagree about exactly which doctors exceed expectations in either direction. At this time, the risk adjustment procedures are not sufficient by themselves to isolate differences attributable solely to quality, nor do they render the process as fair as if all doctors had seen patients of equal complexity.

Finally, focusing the assessment on the process and volume of care rather than on the outcomes may reduce some of the problems associated with complexity. Doctors should take some actions regardless of the patient's severity of illness. For example, the feet of most diabetes patients should be routinely examined, notwithstanding the complexity of their disease. Likewise, volume data are not influenced by patient complexity.

Attribution

Because of increasing complexity in the diagnosis and treatment of patient problems, care now often draws on the skills of doctors with different specialties, nursing professionals, and a variety of other health care providers. Comprehensive and managed care implies efficiently solving patient problems that are beyond the expertise of any single provider. Consequently, patients are often treated by multidisciplinary teams where decision making is shared. This team approach is beneficial for patients but it has significant implications for assessing doctors based on the outcomes of their patients. Specifically, it is difficult to disentangle the effects of the individual doctor from those of the other members of the team.

In a typical assessment situation, the problem of attribution is analogous to giving students a take-home examination where they are free to use their peers or any other of the resources available to them in answering the questions. When the scores are calculated, the degree to which they reflect the true ability of the students is unknown. A relatively weak student might have made an effort to work with a more knowledgeable peer and that peer may suffer by comparison if several other students pooled their resources. Consequently, the results of the test might not be a good predictor of future performance in other settings and would not be useful as a means of comparing a student to another.

Similarly, in a work-based setting, attributing the outcomes of patients to an individual doctor when care was rendered by a multidisciplinary team yields a potentially biased result. An otherwise good doctor might not appear to be so if the other members of the team do not perform well and vice versa. In addition, the performance of a doctor in one team might not be a very good predictor of his or her performance in another team with different membership and different strengths and weaknesses.

One way to address these problems is to include only those outcomes over which doctors have more control. For example, doctors tend to have greater control over the immediate outcomes of simple procedures than they do over the longterm outcomes of chronic illnesses. A focus on these more controllable outcomes as part of the assessment reduces the influence of the health care team and may also lessen the need for sophisticated statistical adjustments. However, it will also result in a view of competence that is biased and perhaps misleading as many challenging and important patient problems do not meet these criteria.

A better way to deal with the problem of attribution might be to focus on process measures that have been shown to correlate with outcomes, rather than the outcomes themselves.[51] For example, with diabetes patients doctors may be assessed on whether they routinely monitor HbA_{1c} or conduct periodic foot and eye examinations. Many of these process measures are directly controlled by the doctor and they are available for a sizeable number of patients. However, good processes do not guarantee good outcomes for patients.

Number of patients

Over the past three decades, one of the more ubiquitous findings in the assessment literature has involved the case specificity of performance. The performance of a doctor with 1 particular patient is not highly correlated with his or her performance with another. Consequently, it takes information from several cases or patients to get a good estimate of competence.[8]

In a traditional assessment setting, this issue is expressed as the reliability or reproducibility of

scores. Unreliable scores used in an educational setting provide an inaccurate picture of an examinee's strengths and weaknesses, leading to misspent educational effort.[52] When they are used as the basis for pass/fail decisions, the result is increased false positive and negative decisions. The single most important influence on reliability is test length because longer tests yield results that are more reliable.

The issues are the same in work-based assessment and initial research suggests that a sizeable number of patients may be needed for good assessment in this area as well. Two studies have examined several outcome-related process measures (e.g. HbA_{1c}, foot and eye examinations, lipids) in the care of diabetes patients.[43,53] There were differences among the various measures in terms of the number of patients required to achieve reliable results, with 1 study recommending 100; both concluded that the number must be sizeable to ensure a reliable assessment of the doctors.

Because of the need for substantial numbers of patients, an assessment based on outcomes or process measures will be limited to only the most common conditions. For other patient problems, a doctor will have too few patients and the results will be unreliable. Consequently, infrequent but important patient problems are not amenable to assessment nor are areas of medicine where change is occurring in the nature of diagnosis or treatment.

CONCLUSION

Traditional assessment has improved significantly over the past 50 years. A number of new testing methods are now in place, the computer is improving both the fidelity and efficiency of examinations, and the psychometric principles on which assessment rests are more sophisticated than ever. However, the growing interest in quality improvement, bolstered by increasing demands for public accountability, has shifted the focus to an assessment of work.

In contrast to traditional assessment, there are no 'methods' for the evaluation of work because 'examiners' are no longer in control of the content. Instead it is a matter of identifying the basis for the judgements (outcomes, process or volume), deciding how the data will be gathered (practice records, administrative databases, diaries/logs or observation), and avoiding threats to validity and reliability (patient mix, patient complexity, attribution and numbers of patients).

Unlike traditional tests, the assessments described in this paper are not focused on the competencies of doctors. Instead they begin with patients and their needs as they encounter the health care system, so the assessments focus on the results of the doctor's actions (i.e. patient outcomes) or specific behaviours (i.e. processes of care). Traditional broad measures of competence may still play an important role, however, especially when there is a need for formative or diagnostic feedback.

Research and development in the assessment of work is just beginning, so much remains to be done. At least in the short-term, outcome-related process measures are preferable as a basis for judgements. Compared to patient outcomes they are more directly in the doctor's control and less susceptible to problems of attribution. Relative to volume, they are a much better indicator of the quality of performance.

The routine use of the electronic medical record may be years away, but it is the best source of information on which to base the assessment of doctors. It makes data collection more feasible and the analysis of data more valid because the richness of the practice records is superior to that of other potential sources of information. Observation will continue to play a key role for the foreseeable future and there is already a body of research that helps to inform its use.

Among the threats to validity and reliability, the development of good risk adjustment procedures should be a priority. These are currently the best and perhaps the only way of levelling the playing field enough to permit meaningful comparisons among doctors and against standards. Another area of priority involves determining the number of patients necessary to achieve good estimates in support of a doctor's assessment. Few studies have been carried out to date and they have focused on limited outcomes, processes and conditions.

Overall, the assessment of doctors' performance at work is in its infancy and much research and development is needed. Nonetheless, it is being used increasingly in programmes of continuous quality improvement and accountability. It is critical that refinements occur quickly to ensure that patients receive the highest quality of care and that doctors are treated fairly and provided with the information they need to guide their professional development.

Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer 2002;647–72.
2 Petrusa ER. Clinical performance assessments. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer 2002;673–710.
3 Cushing A. Assessment of non-cognitive factors. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer 2002;711–56.
4 Wainer H. *Computerized Adaptive Testing: A Primer*. Hillsdale, New Jersey: Lawrence Erlbaum 1990.
5 Green BF. Adaptive testing by computer. In: Ekstrom RB, ed. *Principles of Modern Psychological Measurement*. San Francisco: Jossey-Bass 1983;5–12.
6 Tekian A, McGuire CH, McGaghie WC, eds. *Innovative Simulations for Assessing Professional Competence: From Paper and Pencil to Virtual Reality*. Chicago: University of Illinois 1999.
7 Satava RM. Accomplishments and challenges of surgical simulation. *Surg Endosc* 2001;**15**:232–41.
8 Elstein AS, Shulman LS, Sprafka SA. *Medical Problem-solving: An Analysis of Clinical Reasoning*. Cambridge, Massachusetts: Harvard University Press 1978.
9 Kolen MJ, Brennan RJ. *Test Equating: Methods and Practices*. New York: Springer 1995.
10 Norcini JJ, Guille R. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer 2002;811–34.
11 Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Dordrecht: Kluwer 1985.
12 Brennan RL. *Generalizability Theory*. New York: Springer-Verlag 2001.
13 Norcini JJ, Boulet J. Methodological issues in the use of standardised patients for assessment. *Teach Learn Med* 2003;**15**:293–7.
14 Schuwirth LW, van der Vleuten CP. Changing education, changing assessment, changing research? *Med Educ* 2004;**38**:805–12.
15 Berwick DM, Godfrey AB, Rossener J. *Curing Health Care: New Strategies for Quality Improvement*. San Francisco: Jossey-Bass 1990.
16 Laffel G, Blumenthal D. The case for using industrial quality management science in health care organisations. *JAMA* 1989;**262**:2869–73.
17 Plsek PE. Quality improvement methods in clinical medicine. *Pediatrics* 1999;**103** (Suppl):203–14.
18 Norcini JJ. Psychometric issues in the use of practice performance assessment for physician evaluation. In: Mancall EL, Bashook PG, eds. *Evaluating Residents for Board Certification*. Evanston, Illinois: American Board of Medical Specialties 1998.
19 Norcini JJ. Recertification in the United States. *BMJ* 1999;**319**:1183–5.
20 Norcini JJ. Work-based assessment. *BMJ* 2003;**326**:753–5.
21 Chassin MR. Achieving and sustaining improved quality: lessons from New York State and cardiac surgery. *Health Aff* 2002;**21**:40–51.
22 Mukamel DB, Mushlin AI. Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports. *Med Care* 1998;**36**:945–54.
23 US Preventive Services Task Force. Screening for lipid disorders. recommendations and rationale. *Am J Prev Med* 2001;**20** (3S):73–6.
24 Centers for Disease Control. Recommended childhood and adolescent immunisation schedule – United States. January 2004 http://www.cdc.gov/nip/recs/child-schedule.pdf.
25 Agency for Healthcare Research and Quality. Guide to Clinical Preventive Services. Periodic Updates. 3rd edn. AHRQ Publication No. 04-IP003. Rockville, Maryland: Agency for Healthcare Research and Quality. January 2004. http://www.ahrq.gov/clinic/periodorder.htm.
26 Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Ann Intern Med* 2002;**137**:511–20.
27 Losina E, Barrett J, Mahomed NN, Baron JA, Katz JN. Early failures of total hip replacement: effect of surgeon volume. *Arthritis Rheum* 2004;**50**:1338–43.
28 Norcini JJ, Kimball HR, Lipner RS. Certification and specialisation: do they matter in the outcome of acute myocardial infarction? *Acad Med* 2000;**75**:1193–8.
29 Lorenzoni L, Da Cas R, Aparo UL. The quality of abstracting medical information from the medical record: the impact of training programmes. *Int J Qual Health Care* 1999;**11**:209–13.
30 Gordon NP, Hiatt RA, Lampert DI. Concordance of self-reported data and medical record audit for six cancer screening procedures. *J Natl Cancer Inst* 1993;**85**:566–70.
31 Iezzoni LI. Assessing quality using administrative data. *Ann Intern Med* 1997;**127**:666–74.
32 Larson JL, Look R, Schiffman B. A hand-held computer-based procedure log. *Acad Emerg Med* 2001;**8**:583.
33 Nicolaou DD, Davis GL. A distributed asynchronous resident procedure log for hand-held devices. *Acad Emerg Med* 2001;**8**:583.
34 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65** (Suppl):63–7.
35 Norcini JJ. Peer assessment of competence. *Med Educ* 2003;**37**:539–43.

36 Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004;**79**:16–22.
37 Sturmberg JP. General practice-specific care categories: a method to examine the impact of morbidity on general practice workload. *Fam Pract* 2002;**19**:85–92.
38 Tucker AM, Weiner JP, Honigfeld S, Parton RA. Profiling primary care physician resource use: examining the application of case mix adjustment. *J Ambul Care Manage* 1996;**19**:60–80.
39 Davenport JR, Dennis MS, Warlow CP. Effect of correcting outcome data for case mix: an example from stroke medicine. *BMJ* 1996;**312**:1503–5.
40 Orchard C. Comparing health outcomes. *BMJ* 1994;**308**:1493–6.
41 Hankey GJ, Dennis MS, Slattery JM, Warlow CP. Why is the outcome of transient ischaemic attacks different in different groups of patients? *BMJ* 1993;**306**:1107–11.
42 Kessner DM, Kalk CE, Singer J. Assessing health quality: the case for tracers. *N Engl J Med* 1973;**288**:189–94.
43 Greenfield S, Kaplan SH, Kahn R, Ninomiya J, Griffith JL. Profiling care provided by different groups of physicians: effects of patient case-mix (bias) and physician-level clustering on quality assessment results. *Ann Int Med* 2002;**136**:111–21.
44 Majeed A, Bindman AB, Weiner JP. Use of risk adjustment in setting budgets and measuring performance in primary care. II. Advantages, disadvantages and practicalities. *BMJ* 2001;**323**:607–10.
45 Starfield B, Lemke KW, Bernhardt T, Foldes SS, Forrest CB, Weiner JP. Comorbidity: implications for the importance of primary care in 'case' management. *Ann Fam Med* 2003;**1**:8–14.
46 Tekkis PP, Prytherch DR, Kocher HM, Senapati A, Poloniecki JD, Stamatakis JD, Windsor AC. Development of a dedicated risk-adjustment scoring system for colorectal surgery (colorectal POSSUM). *Br J Surg* 2004;**91**:1174–82.
47 Kuhlthau K, Ferris TG, Iezzoni LI. Risk adjustment for paediatric quality indicators. *Pediatrics* 2004;**113**:210–6.
48 Charlson ME, Pompei P, Ales AL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Clin Epidemiol* 1993;**46**:1075–9.
49 Steen P, Cherney B. Evolution of analytical tools by MediQual Systems, Inc. *Am J Med Qual* 1996;**11**:S15–7.
50 Iezzoni LI. The risks of risk adjustment. *JAMA* 1997;**278**:1600–7.
51 Kremer BK. Physician recertification and outcomes assessment. *Eval Health Prof* 1991;**14**:187–200.
52 Norcini JJ. What should we do about unreliable tests? *Med Educ* 2000;**34**:501–2.
53 Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Mannin WG. The unreliability of individual physician 'report cards' for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;**281**:2098–105.

Received 4 January 2005; accepted for publication 20 January 2005