QUALITY IMPROVEMENT RESEARCH

Research methods used in developing and applying quality indicators in primary care

S M Campbell, J Braspenning, A Hutchinson, M Marshall

Qual Saf Health Care 2002;**11**:358–364

Quality indicators have been developed throughout Europe primarily for use in hospitals, but also increasingly for primary care. Both development and application are important but there has been less research on the application of indicators. Three issues are important when developing or applying indicators: (1) which stakeholder perspective(s) are the indicators intended to reflect; (2) what aspects of health care are being measured; and (3) what evidence is available? The information required to develop quality indicators can be derived using systematic or non-systematic methods. Non-systematic methods such as case studies play an important role but they do not tap in to available evidence. Systematic methods can be based directly on scientific evidence by combining available evidence with expert opinion, or they can be based on clinical guidelines. While it may never be possible to produce an error free measure of quality, measures should adhere, as far as possible, to some fundamental a priori characteristics (acceptability, feasibility, reliability, sensitivity to change, and validity). Adherence to these characteristics will help maximise the effectiveness of quality indicators in quality improvement strategies. It is also necessary to consider what the results of applying indicators tell us about quality of care.

Quality improvement has become a central tenet of health care. It is no longer the preserve of enthusiastic volunteers but part of the daily routine of all those involved in delivering health care, and has become a statutory obligation in many countries. There are numerous reasons why it is important to improve quality of health care, including enhancing the accountability of health practitioners and managers, resource efficiency, identifying and minimising medical errors while maximising the use of effective care and improving outcomes, and aligning care to what users/patients want in addition to what they need.

Quality can be improved without measuring it—for example, by specialist higher educational programmes such as the vocational training scheme for general practice in the UK or guiding care prospectively in the consultation through clinical guidelines.[1 2] Moreover, there are ways of assessing quality without using hard quantitative measures such as quality indicators—for example, peer review, videoing consultations, patient interviews. Measurement, however, plays an important part in improvement[3 4] and helps to promote change.[5] Specific measures may, for example, allow good performance to be rewarded in a fair way and facilitate accountability. For this reason much effort has gone into developing and applying measures of quality over the last few decades. The purpose of this paper is to review methods which seek to develop and apply quality indicators.

See end of article for authors' affiliations

Dr S M Campbell, National Primary Care Research and Development Centre, University of Manchester, Manchester M13 9PL, UK; stephen.campbell@man.ac.uk

DEFINING QUALITY INDICATORS

Indicators are explicitly defined and measurable items which act as building blocks in the assessment of care. They are a statement about the structure, process (interpersonal or clinical), or outcomes of care[6] and are used to generate subsequent review criteria and standards which help to operationalise quality indicators (box 1). Indicators are different from guidelines, review criteria, and standards (box 2). Review criteria retrospectively assess care provided on a case-by-case basis to individuals or populations of patients, indicators relate to care or services provided to patients, and standards refer to the outcome of care specified within these indicators. Standards can be 100%—for example, the National Service Framework for coronary heart disease in the UK has set a standard that all patients with diagnosed coronary heart disease should receive low dose (75 mg) aspirin where clinically appropriate.[7] However, care very rarely meets such absolute standards[8] and, in general, standards should be realistic and set according to local context and patient circumstances.[9 10]

Indicators can measure the frequency with which an event occurred, such as influenza immunisations (activity indicator). However, quality indicators infer a judgement about the quality of care provided.[9] This distinguishes quality indicators from performance indicators,[11] which are statistical devices for monitoring care provided to populations without any necessary inference about quality—for example, they might simply have cost implications. Indicators do not provide definitive answers but *indicate* potential problems that might need addressing, usually manifested by statistical outliers or perceived unacceptable variation in care. Most indicators have been developed to assess/improve care in hospitals but, increasingly, quality measures are being developed for primary care across Europe.

WHAT SHOULD BE MEASURED?

There are three important issues to consider when developing indicators. Firstly, which stakeholder perspective(s) are the indicators intended to reflect? There are different stakeholders of health

Box 1 Definitions of guideline, indicator, review criterion, and standard

Guideline: systematically developed statements to assist practitioner and patient decisions prospectively for specific clinical circumstances; in essence the "right thing to do".[1 2]
Indicator: a measurable element of practice performance for which there is evidence or consensus that it can be used to assess the quality, and hence change in the quality, of care provided.[9]
Review criterion: systematically developed statement relating to a single act of medical care[9] that is so clearly defined it is possible to say whether the element of care occurred or not retrospectively in order to assess the appropriateness of specific healthcare decisions, services, and outcomes.[55 110]
Standard: The level of compliance with a criterion or indicator.[9 77 111] A target standard is set prospectively and stipulates a level of care that providers must strive to meet. An achieved standard is measured retrospectively and details whether a care provider met a predetermined standard.

care (patients, carers, managers, professionals, third party payers).[3 12] It cannot be presumed that one stakeholder's views represent another group's views.[13 14] Different perspectives may need different methods of indicator development, particularly as stakeholders have different perspectives about quality of care. Health professionals tend to focus on professional standards, health outcomes, and efficiency. Patients often relate quality to an understanding attitude, communication skills, and clinical performance. Managers' views are influenced by data on efficiency, patients' satisfaction, accessibility of care and, increasingly, outcomes. Even if the same aspects of care are assessed, the indicator can be valued differently—for example, health professionals and managers will probably value efficiency differently.

Secondly, which aspects of care should be assessed—processes or outcomes of care?[15–18] The ultimate goal of the care given to patients can be expressed as outcome indicators which measure mortality, morbidity, health status, health related quality of life, and patient satisfaction. Examples include medical outcomes,[19] the outcomes utility index,[20] the Computerized Needs Orientated Quality Measurement Evaluation System,[21] and some of the National Performance Frameworks in the UK.[22] Other outcome indicators include user evaluation surveys derived from systematic literature reviews of patient perspectives of health care[23] or outcome indicators developed using focus groups.[24] In this way items included in validated patient surveys such as the General Practice Assessment Survey[25 26] or Europep[27] can be used as quality indicators. One example of such an indicator might be a patient's capacity to get through to practice staff on the telephone. Structural indicators give information on the practice organisation such as personnel, finances, and availability of appointments.[28–31] For example, if a general practice has a car park there should be specified places for disabled parking. There is limited evidence linking structure with outcomes[32] although research has suggested, for example, a link between longer consultations and higher quality clinical care.[21 33 34] Process indicators describe actual medical care such as diagnoses, treatment, referral, and prescribing.[10 35] Since our focus is on quality improvement, our main interest in this paper is on process indicators because improving process has been described as the primary object of quality assessment/improvement.[3 4 16 18 32 36]

Thirdly, in order to develop indicators researchers need information on structure, process or outcome which can be derived in a number of ways using systematic or non-systematic methods. This information is vital to establish the face or content validity of quality measures (box 3).

Box 2 Examples of a guideline, indicator, review criterion, and standard

Guideline recommendation

If a blood pressure reading is raised on one occasion, the patient should be followed up on two further occasions within *x* time.

Indicator

Patients with a blood pressure of more than 160/90 mm Hg should have their blood pressure re-measured within 3 months.

Indicator numerator: Patients with a blood pressure of more than 160/90 mm Hg having had re-measured their blood pressure within 3 months.

Indicator denominator: Patients with a blood pressure of more than 160/90 mm Hg.

Review criterion

If an individual patient's blood pressure was >160/90, was it re-measured within 3 months?

Standard

Target standard: 90% of the patients in a practice with a blood pressure of more than 160/90 mm Hg should have their blood pressure re-measured within 3 months.

Achieved standard: 80% of the patients in a practice with a blood pressure of more than 160/90 mm Hg had their blood pressure re-measured within 3 months.

Box 3 Definitions of acceptability, feasibility, reliability, sensitivity to change, and validity

Development of quality indicators

- Face/content validity: is the indicator underpinned by evidence (content validity) and/or consensus (face validity)? The extent to which indicators accurately represent the concept being assessed (e.g. quality of care for epilepsy).
- Reproducibility: would the same indicators be developed if the same method of development was repeated?

Application of quality indicators

- Acceptability: is the indicator acceptable to both those being assessed and those undertaking the assessment?
- Feasibility: are valid, reliable, and consistent data available and collectable, albeit contained within medical records, health authority datasets or on videotaped consultations?
- Reliability: minimal measurement error, organisations, or practitioners compared with similar organisations or practitioners (comparability), reproducible findings when administered by different raters (inter-rater reliability).
- Sensitivity to change: does the indicator have the capacity to detect changes in quality of care?
- Predictive validity: does the indicator have the capacity for predicting quality of care outcomes?

RESEARCH METHODS FOR THE DEVELOPMENT OF QUALITY INDICATORS

Non-systematic

Non-systematic approaches to developing quality indicators do not tap in to the evidence base of an aspect of health care; they are based on the availability of data and real life critical incidents. This does not mean that they have no useful role in quality assessment/improvement. Examples include quality improvement projects based on one case study.[37] For example, an abortion of a pregnant 13 year old led to a team meeting.[38] Her medical record showed two moments when contraceptives could have been discussed. The response was a special clinic hour for teenagers and the development of a quality

Box 4 What are consensus methods designed to do?

- Enhance decision making,[52] develop policies, and estimate unknown parameters.
- Facilitate the development of quality indicators or review criteria[35 61] where evidence alone is insufficient.
- Synthesise accumulated expert opinion/professional norms.[3]
- Identify, quantify, and subsequently measure areas where there is uncertainty,[47] controversy,[53] or incomplete evidence.[112]

indicator on the administration of lifestyle and risk factors. Other examples include many of the high level indicators used by health authorities[39] and referral rates by general practitioners to specialist services in the UK, as well as many of the VIP indicators of practice development in the Netherlands.[29]

Systematic: evidence based

Where possible, indicators should be based directly upon scientific evidence such as rigorously conducted (trial based) empirical studies.[40–43] The better the evidence, the stronger the benefits of applying the indicators in terms of reduced morbidity and mortality or improved quality of care. For example, patients with confirmed coronary artery disease should be prescribed aspirin, unless contraindicated, as there is evidence that aspirin is associated with improved health benefits in patients with coronary heart disease, although the evidence on the exact dose is unclear. McColl and colleagues have developed sets of evidence-based indicators for use by primary care organisations in the UK based on available data.[44]

Systematic: evidence combined with consensus

There are, however, many grey areas of health care for which the scientific evidence base is limited,[45] especially within the generalist and holistic environment of general practice. This necessitates using an extended family of evidence to develop quality indicators, including utilising expert opinion.[42 46 47] However, experts often disagree on the interpretation of evidence and rigorous and reproducible methods are needed to assess the level of agreement; in particular, combining expert opinion with available evidence using consensus techniques to assess aspects of care for which evidence alone is insufficient, absent, or methodologically weak.[9 41 48] The idea of harvesting professional opinion regarding professional norms of practice to develop quality measures is not new.[3]

Box 4 shows that there are a variety of reasons for developing quality indicators using consensus methods. They also allow a wider proportion of aspects of quality of care to be assessed and thus improved than if indicators were based solely on evidence. Quality indicators abound for preventive care, are patchy for chronic care, and almost absent for acute care in general practice.[49]

Consensus techniques are group facilitation techniques which explore the level of consensus among a group of experts by synthesising and clarifying expert opinion in order to derive a consensus opinion from a group with individual opinions combined into a refined aggregated opinion. Group judgements of professional opinion are preferable to individual practitioner judgements because they are more consistent; individual judgements are more prone to personal bias and lack of reproducibility. Recent examples include quality indicators for common conditions,[10] research on the necessity of process indicators for quality improvement,[50] and a practice visit tool to augment quality improvement.[29]

There are a number of techniques including the Delphi technique[51–53] and the RAND appropriateness method[54] which have been discussed elsewhere,[41] and guideline driven indicators using an iterated consensus rating procedure.[55] The nominal group technique[56] is also used in which a group of experts is asked to generate and prioritise ideas but it is not itself a consensus technique.[41] However, the nominal group technique, supported by postal Delphi, has been used to produce, for example, a national clinical practice guideline in the UK[57] and prescribing indicators.[58]

Delphi technique

The Delphi technique is a structured interactive method involving repetitive administration of anonymous questionnaires, usually across two or three postal rounds. Face to face meetings are not usually a feature. The main stages include: identifying a research problem, developing questionnaire statements to rate, selecting appropriate panellists, conducting anonymous iterative postal questionnaire rounds, feeding back results (statistical, qualitative, or both) between rounds, and summarising and feeding back the findings.

The approach enables a large group to be consulted from a geographically dispersed population. For example, Shield[59] used 11 panels composed of patients, carers, health managers, and health professionals to rate quality indicators of primary mental health care. Optimal size has not been established and research has been published based on samples ranging from 4 to 3000.

The Delphi procedure permits the evaluation of large numbers of scenarios in a short time period.[60] The avoidance of face to face interaction between group members can prevent individuals feeling intimidated and opinions can be expressed away from peer group pressure. However, the process of providing group and, particularly, individual feedback can be very resource intensive. Moreover, the absence of any face to face panel discussion prohibits the opportunity to debate potentially different viewpoints. There is limited evidence of the validity of quality measures derived using the Delphi technique.[41 52] The Delphi procedure has been used to develop prescribing indicators,[61] managerial indicators,[62] indicators of patient and general practitioner perspectives of chronic illness,[23] indicators for cardiovascular disease,[63] and key attributes of a general practice trainer.[64] The Delphi technique has therefore been used to generate indicators for more than just clinical care.

RAND appropriateness method

This method is a formal group judgement process which systematically and quantitatively combines expert opinion and scientific (systematic literature review) evidence by asking panellists to rate, discuss, and then re-rate indicators. It is the only systematic method of combining expert opinion and evidence.[65] It also incorporates a rating of the feasibility of collecting data, a key characteristic in the application of indicators as discussed below. The main stages include selection of the condition(s) to be assessed, a systematic literature review of the available evidence, generation of preliminary indicators to be rated, selection of expert panels, first round postal survey where panellists are asked to read the accompanying evidence and rate the preliminary indicators, a face to face panel meeting where panellists discuss each indicator in turn, analyses of final ratings, and development of recommended indicators/criteria.[48] The method has been the subject of a number of critiques.[48 65–68]

The RAND method has been used most often to develop appropriateness criteria for clinical interventions in the US[69 70] such as coronary angioplasty or for developing quality indicators for assessing care of vulnerable elderly patients.[71] It has also been used in the UK,[72–74] including the development of review criteria for angina, asthma and diabetes[35 75] and for 19 common conditions including acute, chronic and preventive care.[10]

The strengths of the RAND method are that panellists meet so discussions can take place, no indicators are discarded between rounds so no potential information is lost and, unlike the standard Delphi technique, panellists are sent a copy of the systematic literature review in addition to the catalogue of indicators. This increases the opportunities for panel members to ground their opinions in the scientific evidence. Research

Table 1 Guideline driven indicators developed using an iterated consensus rating procedure

	Aim	Undertaken by	Criteria used
Round 1: Pre-selection	Selecting key recommendations	Small group of quality indicators developers (1–3 persons)	Outcome of care: • Patients' health (morbidity, mortality, health status) • Cost
Round 2: Rating and adding	Rating and adding key recommendations	Expert panel (8–10 persons)	• Patients' health • Cost • Sensitivity to change • Availability of data
Round 3: Reliability	Determining inter- and intra-rater reliability	Expert panel for the rating Research team for the analyses	• Kappa, rho
Round 4: Potential indicators	Getting set of potential indicators	Research team	• Cut off score: mean above mid of rating scale • Agreement among 80% of the panel members
Round 5: Reflection	Acceptability of indicators	Research team Laymen professionals	• Face validity

has also shown that using a higher cut off point for determining consensus within a panel (an overall panel median rating of 8 out of 9) enhances the reproducibility (box 3) of the ratings if a different set of panellists rated the indicators.[76] Shekelle and colleagues found that, while agreement between panels was weak, in terms of kappa values they had greater reliability than many widely accepted clinical procedures such as reading of mammograms.[48]

However, the panels inevitably have to be smaller than the Delphi panels for practical reasons, users/patients are rarely involved, the implications of costs are not considered in ratings, and indicators have been limited to clinical care. Moreover, the face to face nature of the discussion can lead to potential intimidation if there are dominant personalities, although each panellists' ratings carry equal weight irrespective of how much/little they contribute to the discussion.

Systematic: guideline driven indicators

Indicators can be based on clinical guidelines.[55 77–79] Such indicators for general practice have been developed and disseminated widely in the NHS in the UK for four important clinical conditions (diabetes, coronary heart disease, asthma, and depression),[80] using methods proposed by AHCPR.[55] Review criteria were derived from at least one clinical guideline which met a set of quality standards, using structured questions and feedback to test the face and content validity—as well as the feasibility—of the criteria with a panel of over 60 general practitioners.

Hadorn and colleagues[81] described how 34 recommendations in a guideline on heart failure were translated into eight review criteria. Because review criteria must be specific enough to assure the reliability and validity of retrospective review, they used two selection criteria to guide whether each recommendation based criterion should be retained in the final selection—importance to quality of care and feasibility of monitoring. They demonstrated some important aspects of criteria development from guidelines, in particular the need to be very detailed and specific in the criterion, even though the guideline recommendation is less specific and deemed adequate.

Review criteria derived directly from a clinical practice guideline are now part of NHS policy in England and Wales through the work of the National Institute of Clinical Excellence (NICE). Each published summary clinical guideline is accompanied by a set of review criteria which are intended to be used by clinical teams, and the results are externally assessed by the Commission for Health Improvement—for example, in relation to type 2 diabetes.[82] These NICE criteria were developed using an iterated consensus rating procedure similar to that used frequently by the Dutch College of General Practitioners—for example, for back pain and the management of stroke treatment in hospitals. The prominent method in the Netherlands is an iterated consensus rating procedure which seeks to develop indicators based on the impact of guideline recommendations on the outcomes of care (table 1).[55 79] Developers reflect critically on the acceptability of developed sets in conjunction with a group of lay professionals. The method has evolved within the last decade. Some initial studies assessed the performance of the general practitioner on, for example, threatened miscarriage, asthma and chronic obstructive pulmonary disease where the indicator development was limited to the first round of the procedure.[83 84] Other studies used larger panels to assess key recommendations.[85–87] More recent projects have completed all five rounds—for example, a study in which quality indicators were selected for all 70 guidelines developed by the Dutch College of General Practitioners[55] or a study on the management of stroke in hospital.[79]

FACTORS INFLUENCING THE DEVELOPMENT OF QUALITY INDICATORS USING A CONSENSUS TECHNIQUE

Many factors influence ratings in a consensus method,[41] especially group composition as groups composed of different stakeholders rating the same statements produce different ratings.[2 66 73 88 89] For example, group members who use, or are familiar with, the procedures being rated are more likely to rate them higher.[69 70 89 90] Moreover, panel members from different disciplines make systematically different judgements and feedback from mixed disciplines may influence ratings. For example, a Delphi composed equally of health physicians and managers found that the physicians who had overall feedback, including that of the managers, rated indicators higher than the physicians who had physician only feedback, whereas managers with combined feedback rated lower than managers with manager only feedback.[88]

Ongoing work has provided qualitative evidence of factors which influence individual panellists' ratings in a consensus technique rating aspects of the quality of primary mental health care in a two round postal Delphi.[59] This research used in depth qualitative interviews with panellists from patient, managerial, and professional panels to identify factors which had influenced panellists' ratings. It concluded that many factors influenced the ratings of the different stakeholder groups (box 5).

RESEARCH METHODS ON THE APPLICATION OF INDICATORS

Measures derived using expert panels and guidelines have high face validity and those based on rigorous evidence possess high content validity. However, this should be a minimum prerequisite for any quality measure and subsequent developmental work is required to provide empirical evidence, as far as possible, of acceptability, feasibility, reliability, sensitivity to change, and predictive validity (box 3).[6 68 91 92]

Box 5 Factors influencing indicators rated valid in a Delphi technique[41 59]

- Composition of the panel
- Inclusion of patient derived (focus groups) indicators
- Inclusion of indicators based on "grey" literature
- Inclusion of multiple stakeholders (e.g. patients, carers, managers, health professionals)
- Characteristics of individual panellists (e.g. political perspective, familiarity with research)
- Rating process (e.g. 9 point scale, feedback used)
- Panellists' experience and expectations of the care provision being rated
- Panellists' perspective of the model of care provision
- Panellists' perspective of their locus of control to influence care

Acceptability

The acceptability of the data collected using a measure will depend upon the extent to which the findings are acceptable to both those being assessed and those undertaking the assessment. For example, the iterated consensus rating procedure consults lay professionals as to the acceptability of indicators (table 1). Campbell and colleagues conducted a quality assessment in 60 general practices in England but only used quality indicators rated acceptable and valid by the nurses and doctors working in the practices.[75]

Feasibility

Information about the quality of services is often driven by data availability rather than by epidemiological and clinical considerations.[93] Quality measurement cannot be achieved without accurate and consistent information systems.[15 94] Current administrative data, both at the macro (health authority or "large organisation") and micro (individual medical records) levels, are constrained by inconsistent and often unreliable data.[95–98] Medical records are a poor vehicle for collecting data on preventive care and the recording of symptoms.[99–101]

In addition, aspects of care being assessed by quality indicators must relate to enough patients to make comparing data feasible. For example, a clinical audit of angina care excluded 10 criteria rated necessary by an expert panel to provide quality of care[35] because they related to less than 1% of a sample of over 1000 patients in 60 general practices in England.[75]

Reliability

Indicators should be used to compare organisations/practitioners with similar organisations/practitioners, or confounding factors such as socioeconomic and demographic factors, as well as factors outside the control of practitioners, should be taken into account (that is, compare like with like or risk/case mix adjust). This is because the environment in which an organisation operates affects the care provided. Examples include admission rates or surgery rates. Indicators must also have explicit exclusion and inclusion criteria for applying the indicator to patients—for example, age ranges, co-morbidities, case mix, and clinical diagnoses.

The inter-rater reliability of an indicator can also be tested when applying indicators. For example, in a study of over 1000 patients with diabetes two raters abstracted data separately (but on the same day) for 7.5% of all patient records and found that five criteria out of 31 developed using an expert panel were excluded from analyses due to poor agreement.[75]

Sensitivity to change

Quality measures must be capable of detecting changes in quality of care[17] in order to discriminate between and within subjects.[91] This is an important and often forgotten dimension of Lawrence's definition of a quality indicator.[9]

Validity

There has been little methodological scrutiny of the validity of consensus methods.[42 46 92 102] The Delphi technique[103] and the RAND method[16 104] have both been criticised for a lack of evidence of validity. While the issue has received more attention in recent years,[6 16 36] there is little evidence for the validity of the Delphi method in developing quality indicators.

Content validity of indicators generated using consensus techniques

Content validity in this context refers to whether any indications were rated by panels contrary to known results from randomised controlled trials. There is evidence for the content validity of indicators derived using the RAND method.[48 105]

Predictive validity

There is evidence of the predictive validity of indicators developed using the RAND method.[48 106 107] For example, Kravitz and colleagues studied a cohort of persons who had undergone coronary angiography. Patients were retrospectively classified as to whether coronary revascularisation was "necessary" or "not necessary" according to the review criteria, and outcomes at year 1 were measured. Patients meeting the "necessary" criteria for coronary revascularisation who did not receive it were twice as likely to have died at 1 year as those who did receive "necessary" revascularisation. Hemingway *et al*[74] found substantial underuse of coronary revascularisation among UK patients who were considered appropriate for these procedures and underuse was associated with adverse clinical outcomes on the basis of the ratings of an expert panel.

USING DATA GENERATED BY APPLYING QUALITY INDICATORS

Data generated using quality indicators can be used for a variety of purposes—for example, to monitor, reward, penalise, or compare care provision (perhaps using league tables or public release of data) or as part of a quality improvement strategy. Simply measuring something will not automatically improve it. Indicators must be used within coherent systems based approaches to quality improvement.[108 109] The interpretation and usage of such data is more of a political or resource issue than a methodological or conceptual one.

The provenance of the indicators is important when applying them. Indicators derived from informal consensus procedures with little evidence underlying them might be useful as educational guidelines. However, the best indicators for public disclosure, for use in league tables, or for attaching financial incentives are those based solely on scientific evidence, for which the implications of applying the indicator and any relative judgements that are be inferred about the results can be confidently predicted. Indicators derived from consensus methods which systematically combine evidence and opinion may also be disclosed, but perhaps with more provisos. Indicators developed by well respected experts using a systematic method might also have high credibility when used for professional development.

CONCLUSION

It may never be possible to produce an error free measure of quality, but measures should adhere, as far as possible, to some fundamental a priori characteristics in their development (face/content validity) and application (acceptability, feasibility, reliability, sensitivity to change, predictive validity). Adherence to these characteristics will help maximise the effectiveness of quality indicators in quality improvement strategies. This is most likely to be achieved when they are derived from rigorous scientific evidence. However, evidence in health care is often absent. We believe that using consensus techniques—which systematically combine evidence and opinion—and guideline driven approaches facilitates quality

Key messages

- Most quality indicators have been developed in hospitals but they are increasingly being developed for primary care in Europe and the USA.
- Most research has focused on the development rather than the application of indicators.
- Quality indicators should be based on rigorous scientific evidence if possible. However, evidence in health care is often absent, necessitating the use of other methods of development including consensus techniques (such as the Delphi technique and the RAND appropriateness method) which combine expert opinion and available evidence and indicators based on clinical guidelines.
- While it may never be possible to produce an error free measure of quality, measures should adhere, as far as possible, to some fundamental a priori characteristics—namely, acceptability, feasibility, reliability, sensitivity to change, and validity.
- The way in which indicators are applied is as important as the method of development.

improvement. They allow a significantly broader range of aspects of care to be assessed and improved than would be the case if quality indicators were restricted to scientific evidence.

It is important that such methods of development continuously improve and seek to incorporate advances in the evidence base of health care. However, it may be that research has reached a peak in developing indicators. There is much less research on the application of indicators and their reliability, validity, and effectiveness in quality improvement strategies, how indicators can be used to improve care, and how professionals/service users can be helped to be more engaged with the development and use of indicators. Introducing strategies for quality improvement based on quality indicators does not make them effective and successful without understanding the factors that are required to underpin their development and to facilitate their transference between settings and countries.

Authors' affiliations

S M Campbell, M Marshall, National Primary Care Research and Development Centre, University of Manchester, Manchester M13 9PL, UK
A Hutchinson, University of Sheffield, Section of Public Health, ScHARR, Sheffield S1 4DA, UK
J Braspenning, UMC St Radboud, WOK, Centre for Quality of Care Research, Postbus 9101, 6500 HB Nijmegen, The Netherlands

REFERENCES

1 **Forrest D**, Hoskins A, Hussey. Clinical guidelines and their implementation. *Postgrad Med J* 1996;**72**:19–22.
2 **Grimshaw JM**, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *Lancet* 1993;**342**:1317–22.
3 **Donabedian A**. *Explorations in quality assessment and monitoring. Volume 1: The definition of quality and approaches to its assessment*. Ann Arbor, Michigan: Health Administration Press, 1980.
4 **Irvine D**. *Managing for quality in general practice*. London: King's Fund Centre, 1990.
5 **Juran JM**. *Juran on planning for quality*. New York: Free Press, 1988.
6 **McGlynn EA**, Asch SM. Developing a clinical performance measure. *Am J Prevent Med* 1998;**14**:14–21.
7 **Department of Health**. *A National Service Framework for coronary heart disease*. London: Department of Health, 2000.
8 **Seddon ME**, Marshall MN, Campbell SM, *et al*. Systematic review of studies of clinical care in general practice in the United Kingdom, Australia and New Zealand. *Qual Health Care* 2001;**10**:152–8.
9 **Lawrence M**, Olesen F, *et al*. Indicators of quality health care. *Eur J Gen Pract* 1997;**3**:103–8.
10 **Marshall M**, Campbell SM. Introduction to quality assessment in general practice. In: Marshall M, Campbell SM, Hacker J, Roland MO, eds. *Quality indicators for general practice: a practical guide for health professionals and managers*. London: Royal Society of Medicine, 2002: 1–6.
11 **Buck D**, Godfrey C, Morgan A. *Performance indicators and health promotion targets*. Discussion paper 150. York: Centre for Health Economics, University of York, 1996.
12 **Ovretveit J**. *Health service quality: an introduction to quality methods for health services*. Oxford: Blackwell Scientific Publications, 1992.
13 **McGlynn EA**. Six challenges in measuring the quality of health care. *Health Aff* 1997;**16**:7–21.
14 **Joss R**, Kogan M. *Advancing quality: total quality management in the National Health Service*. Buckingham: Open University Press, 1995.
15 **Davies HTO**, Crombie IK. Assessing the quality of care. *BMJ* 1995;**311**:766.
16 **Eddy DM**. Performance measurement: problems and solutions. *Health Aff* 1998;**17**:7–26.
17 **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.
18 **Palmer RH**. Process-based measures of quality: the need for detailed clinical data in large health care databases. *Ann Intern Med* 1997;**127**:733–8.
19 **Tarlov AR**, Ware JE, Greenfield S, *et al*. The Medical Outcomes Study: an application of methods for monitoring the results of medical care. *JAMA* 1989;**262**:925–30.
20 **McGlynn EA**. The outcomes utility index: will outcomes data tell us what we want to know? *Int J Qual Health Care* 1998;**10**:485–90.
21 **Agency for Healthcare Research and Quality**. *Computerized needs: oriented quality measurement evaluation system*. Rockville: Agency for Healthcare Research and Quality, 1999 (www.ahrq.gov/qual/conqix.htm).
22 **NHS Executive**. *Quality and performance in the NHS: high level performance indicators*. London: Department of Health, 1999.
23 **Roland MO**, Holden J, Campbell SM. *Quality assessment for general practice: supporting clinical governance in primary care groups*. Manchester: National Primary Care Research and Development Centre, 1998.
24 **Wensing M**, Jung HP, Mainz J, *et al*. A systematic review of the literature on patient priorities for general practice care. Part 1: Description of the research domain. *Soc Sci Med* 1998;**47**:1573–88.
25 **Campbell SM**, Hann M, Hacker J, *et al*. Identifying predictors of high quality care in English general practice: an observational study. *BMJ* 2001;**323**:784–7.
26 **Ramsay J**, Campbell JL, Schroter S, *et al*. The General Practice Assessment Survey (GPAS): tests of data quality and measurement properties. *Fam Pract* 2000;**17**:372–9.
27 **Wensing M**, Vedsted P, Kersnik J, *et al*. Patient satisfaction with availability of general practice: an international comparison. *Int J Qual Health Care* 2002;**14**:111–8.
28 **National Committee for Quality Assurance**. *Narrative: What's in it and why it matters*. Volume 1. HEDIS 3.0/1998. Washington: National Committee for Quality Assurance, 1998.
29 **Van den Hombergh P**, Grol R, van den Hoogen HJ, *et al*. Practice visits as a tool in quality improvement: mutual visits and feedback by peers compared with visits and feedback by non-physician observers. *Qual Health Care* 1999 **8**:161–6.
30 **American Academy of Family Physicians**. *The Family Practice Management Practice Self-Test*. 2001 (available at: www.aafp.org/fpm/20010200/41thef.html)
31 **Royal Australian College of General Practitioners**. *Standards for general practice*. Royal Australian College of General Practitioners, 2000.
32 **Brook RH**, McGlynn EA, Cleary PD. Measuring quality of care. *N Engl J Med* 1996;**335**:966–70.
33 **Howie JG**, Heaney DJ, Maxwell M. Measuring quality in general practice. Pilot study of a needs, process and outcome measure. *Occasional Paper of the Royal College of General Practitioners* 1997;**75**:1–32.
34 **Wilson A**, Childs S. *Systematic review on consultation length in general practice*. A report to the Scientific Foundation Board of the RCGP, University of Leicester, Leicester, 2001.
35 **Campbell SM**, Roland MO, Shekelle PG, *et al*. Development of review criteria for assessing the quality of management of stable angina, adult asthma and non-insulin dependent diabetes in general practice. *Qual Health Care* 1999;**8**:6–15.
36 **Brook RH**, McGlynn EA, Shekelle PG. Defining and measuring quality of care: a perspective from US researchers. *Int J Qual Health Care* 2000;**12**:281–95.
37 **Pringle M**. Preventing ischaemic heart disease in one general practice: from one patient, through clinical audit, needs assessment, and commissioning into quality improvement. *BMJ* 1998;**317**:1120–4.
38 **Pringle M**. Clinical governance in primary care. Participating in clinical governance. *BMJ* 2000;**321**:737–40.
39 **NHS Executive**. *Quality and performance in the NHS: high level performance indicators*. London: Department of Health, 1999.
40 **Hearnshaw HM**, Harker RM, Cheater FM, *et al*. Expert consensus on the desirable characteristics of review criteria for improvement of health quality. *Qual Health Care* 2001;**10**:173–8.
41 **Campbell SM**, Cantrill JA. Consensus methods in prescribing research. *J Clin Pharm Ther* 2001;**26**:5–14.
42 **Murphy MK**, Black NA, Lamping DL, *et al*. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;**2**(3).
43 **Baker R**, Fraser RC. Is ownership more important than the scientific credibility of audit protocols? A survey of medical audit advisory groups. *Fam Pract* 1997;**14**:107–11.
44 **McColl A**, Roderick P, Gabbay J, *et al*. Performance indicators for primary care groups: an evidence-based approach. *BMJ* 1998;**317**:1354–60.

45 **Naylor CD**. Grey zones in clinical practice: some limits to evidence based medicine. *Lancet* 1995;**345**:840–2.
46 **Black N**, Murphy M, Lamping D, *et al*. Consensus development methods: a review of best practice in creating clinical guidelines. *J Health Serv Res Policy* 1999;**4**:236–48.
47 **Jones JJ**, Hunter D. Consensus methods for medical and health services research. *BMJ* 1995;**311**:376–80.
48 **Shekelle PG**, Kahan JP, Bernstein SJ, *et al*. The reproducibility of a method to identify the overuse and underuse of procedures. *N Engl J Med* 1998;**338**:1888–95.
49 **Campbell SM**, Roland MO, Buetow S. Defining quality of care. *Soc Sci Med* 2000;**51**:1611–25.
50 **Ibrahim JE**. Performance indicators form all perspectives. *Int J Qual Health Care* 2001;**13**:431–2.
51 **Linstone HA**, Turoff M. *The Delphi survey. Method, techniques and applications*. Reading, Massachusetts: Addison-Wesley, 1975.
52 **Hasson F**, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Advan Nurs* 2000;**32**:1008–15.
53 **Fink A**, Kosecoff J, Chassin M, *et al*. Consensus methods: characteristics and guidelines for use. *Am J Public Health* 1984;**74**:979–83.
54 **Brook RH**, Chassin MR, Fink A, *et al*. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care* 1986;**2**:53–63.
55 **Braspenning J**, Drijver R, Schiere AM. *Quality indicators for general practice* (in Dutch). Nijmegen/Utrecht: Centre for Quality of Care Research/ Dutch College of General Practitioners, 2001.
56 **Delbecq AL**, Van de Ven AH, Gustafson D. *Group techniques for programme planning: a guide to nominal group and Delphi processes*. Glenview, Illinois: Scott, Foresman & Company, 1975.
57 **Department of Health**. *Treatment choice in the psychological therapies and counselling*. London: Department of Health, 2001.
58 **Cantrill JA**, Sibbald B, Buetow S. Indicators of the appropriateness of long term prescribing in general practice in the United Kingdom: consensus development, face and content validity, feasibility and reliability. *Qual Health Care* 1998;**7**:130–5.
59 **Shield TL**. *Quality indicators for mental health care in primary care*. Personal correspondence, NPCRDC, Manchester, 2002.
60 **Rockwell MA**. The Delphi procedure: knowledge from goat viscera? *N Engl J Med* 1973;**288**:1298–9.
61 **Campbell SM**, Cantrill JA, Richards D. Prescribing indicators for UK general practice: Delphi consultation study. *BMJ* 2000;**321**:1–5.
62 **Campbell SM**, Roland MO, Quayle JA, *et al*. Quality indicators for general practice. Which ones can general practitioners and health authority managers agree are important and how useful are they? *J Public Health Med* 1998;**20**:414–21.
63 **Normand SL**, McNeil BJ, Peterson LE, *et al*. Eliciting expert opinion using the Delphi technique: identifying performance indicators for cardiovascular disease. *Int J Qual Health Care* 1998;**10**:247–60.
64 **Munro N**, Hornung RI, McAteer S. What are the key attributes to of a good general practice trainer? A Delphi study. *Educ Gen Pract* 1998;**9**:263–70.
65 **Naylor CD**. What is appropriate care? *N Engl J Med* 1998;**338**:1918–20.
66 **Hicks NR**. Some observations on attempts to measure appropriateness of care. *BMJ* 1994;**309**:730–3.
67 **Ayanian JZ**, Landrum MB, Normand SLT, *et al*. Rating the appropriateness of coronary angiography – do practicing physicians agree with an expert panel and with each other? *N Engl J Med* 1998;**338**:1896–904.
68 **Campbell SM**. *Defining, measuring and assessing quality of care in general practice*. PhD Thesis, University of Manchester, Manchester, 2002.
69 **Leape LL**, Hilborne LH, Schwartz JS, *et al*. The appropriateness of coronary artery bypass graft surgery in academic medical centres. *Ann Intern Med* 1996;**125**:8–18.
70 **Kahn KL**, Rogers WH, Rubenstein LV, *et al*. Measuring quality of care with explicit process criteria before and after implementation of a DRG-based prospective payment system. *JAMA* 1990;**264**:1969–73.
71 **Shekelle PG**, MacLean CH, Morton SC, *et al*. Assessing care of vulnerable elders: methods for developing quality indicators. *Ann Intern Med* 2001;**135**:647–52.
72 **Gray D**, Hampton JR, Bernstein SJ, *et al*. Audit of coronary angiography and bypass surgery. *Lancet* 1990;**335**:1317–20.
73 **Scott EA**, Black N. Appropriateness of cholecystectomy in the United Kingdom: a consensus panel approach. *Gut* 1991;**32**:1066–70.
74 **Hemingway H**, Crook AM, Feder G, *et al*. Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization. *N Engl J Med* 2001;**344**:645–54.
75 **Campbell SM**, Hann M, Hacker J, *et al*. Quality assessment for three common conditions in primary care: validity and reliability of review criteria developed by expert panels for angina, asthma and type 2 diabetes. *Qual Saf Health Care* 2002;**11**:125–30.
76 **Shekelle PG**, Kahan JP, Park RE, *et al*. Assessing appropriateness by expert panels: how reliable? *J Gen Intern Med* 1996;**10**:81.
77 **Eccles M**, Clapp Z, Grimshaw J, *et al*. North of England evidence based guidelines development project: methods of guideline development. *BMJ* 1996;**312**:760–2.
78 **Agency for Healthcare Research and Quality**. *Using clinical practice guidelines to evaluate quality of care. Volume 2, Methods*. Rockville: Agency for Healthcare Research and Quality, 1995.
79 **CBO Quality Institute for Health Care**. *Handbook. Development of indicators on vidence-based guidelines* (in Dutch). Utrecht: Quality Institute for Health Care, 2002.
80 **Hutchinson A**, Anderson JP, McIntosh A, *et al*. *Evidence based review criteria for coronary heart disease*. Sheffield: Royal College of General Practitioners Effective Clinical Practice Unit, University of Sheffield, 2000.
81 **Hadorn DC**, Baker DW, Kamberg CJ, *et al*. Phase II of the AHCPR-sponsored heart failure guideline: translating practice recommendations into review criteria. *J Qual Improve* 1996;**22**:266–75.
82 **National Institute for Clinical Excellence**. *Management of type 2 diabetes: renal disease-prevention and early management*. London: National Institute for Clinical Excellence, 2002.
83 **Smeele IJ**, Grol RP, van Schayck CP, *et al*. Can small group education and peer review improve care for patients with asthma/chronic obstructive pulmonary disease? *Qual Health Care* 1999;**8**:92–8.
84 **Grol R**, Dalhuijsen J, Thomas S, *et al*. Attributes of clinical guidelines that influence use of guidelines in genral practice: observational study. *BMJ* 1998;**317**:858–61.
85 **Spies TH**, Mokkink HGA. *Using guidelines in clinical practice* (in Dutch). Nijmegen/Utrecht: Centre for Quality of Care Research/ Dutch College of General Practitioners, 1999.
86 **Schers H**, Braspenning J, Drijver R, *et al*. Low back pain in general practice: reported manegement and reasons for not adhering to the guidelines in the Netherlands. *Br J Gen Pract* 2000;**50**:640–4.
87 **Frijling BD**, Spies TH, Lobo CM, *et al*. Blood pressure control in treated hypertensive patients: clinical performance of general practitioners. *Br J Gen Pract* 2001;**51**:9–14.
88 **Campbell SM**, Hann M, Roland MO, *et al*. The effect of panel membership and feedback on ratings in a two-round Delphi survey. *Med Care* 1999;**37**:964–8.
89 **Coulter I**, Adams A, Shekelle P. Impact of varying panel membership on ratings of appropriateness in consensus panels: a composition of a multi- and single disciplinary panel. *Health Serv Res* 1995;**30**:577–91.
90 **Fraser GM**, Pilpel D, Kosecoff J, *et al*. Effect of panel composition on appropriateness ratings. *Int J Qual Health Care* 1994;**6**:251–5.
91 **Streiner DL**, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford Medical Publications, 1995.
92 **Huff ED**. Comprehensive reliability assessment and comparison of quality indicators and their components. *J Clin Epidemiol* 1997;**50**:1395–404.
93 **Siu AL**, McGlynn EA, Morgenstern H, *et al*. Choosing quality of care measures based on the expected impact of improved care on health. *Health Serv Res* 1992;**27**:619–50.
94 **Thomson R**, Lally J. Clinical indicators: do we know what we're doing? *Qual Health Care* 1998;**7**:122.
95 **Enthoven AC**. A promising start, but fundamental reform is needed. *BMJ* 2000;**320**:1329–21.
96 **Wilkinson EK**, McColl A, Exworthy M, *et al*. Reactions to the use of evidence-based performance indicators in primary care: a qualitative study. *Qual Health Care* 2000;**9**:166–74.
97 **Baker R**. Managing quality in primary health care: the need for valid information about performance. *Qual Health Care* 2000;**9**:83.
98 **Craddock J**, Young A, Sullivan G. The accuracy of medical record documentation in schizophrenia. *J Behav Health Serv Res* 2001;**28**:456–66.
99 **Wyatt JC**, Wright P. Design should help use of patients' data. *Lancet* 1998;**352**:1375–8.
100 **Wu L**, Ashton CM. Chart review: a need for reappraisal. *Evaluating Health Professionals* 1997;**20**:146–63.
101 **Luck J**, Peabody JW, Dresselhaus TR, *et al*. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *Am J Med* 2000;**108**:642–9.
102 **Salzer MS**, Nixon CT, Schut LJA, *et al*. Validating quality indicators: quality as relationship between structure, process and outcome. *Evaluation Rev* 1997;**21**:292–309.
103 **Kahn KL**, Park RE, Vennes J, *et al*. Assigning appropriateness ratings for diagnostic upper gastrointestinal endoscopy using two different approaches. *Med Care* 1992, **30**:1016–28.
104 **Phelps CE**. The methodologic foundations of studies of the appropriateness of medical care. *N Engl J Med* 1993;**329**:1241–5.
105 **Merrick NJ**, Fink A, Park RE, *et al*. Derivation of clinical indications for carotid endarterectomy by an expert panel. *Am J Public Health* 1987;**77**:187–90.
106 **Kravitz RL**, Laouri M, Kahan JP, *et al*. Validity of criteria used for detecting underuse of coronary revascularization. *JAMA* 1995;**274**:632–8.
107 **Selby JV**, Fireman BH, Lundstrom. Variation among hospitals in coronary-angiography practices and outcomes after myocardial infarction in a large health maintenance organisation. *N Engl J Med* 1996;**335**:1888–96.
108 **Ferlie EB**, Shortell SM. Improving the quality of health care in the United Kingdom and the United States: a framework for change. *Milbank Quarterly* 2001;**79**:281–315.
109 **Campbell SM**, Sweeney GM. The role of clinical governance as a strategy for quality improvement in primary care. *Br J Gen Pract* 2002 (in press).
110 **Donabedian A**. *Explorations in quality assessment and monitoring. Volume 2. The criteria and standards of quality*. Ann Arbor, Michigan: Health Administration Press, 1982.
111 **Donabedian A**. The quality of medical care. *Science* 1978;**2000**:856–64.
112 **Lomas J**, Anderson G, Enkin M, *et al*. The role of evidence in the consensus process. Results from a Canadian consensus exercise. *JAMA* 1988;**259**:3001–5.