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

S M Campbell, M Hann, J Hacker, A Durie, A Thapar, M O Roland

Original Article

Quality of care has been defined by a number of researchers and there are a variety of methods available for its measurement including clinical indicators or review criteria. The UK government has developed sets of clinical indicators for the National Health Service (NHS), most of which are focused on secondary care or public health, but there is an increasing focus on primary care indicators.

McGlynn set six challenges for measuring quality of care, including establishing credible criteria. While it may never be possible to produce an error-free measure of quality of care, measures should be based on characteristics of best practice such as validity, reliability, and transparency and include instruments created for use with medical records.

Previous work by two of the authors (SC, MR) used expert panels to develop evidence-based review criteria for angina, asthma, and type 2 diabetes. These panels developed 42 criteria for angina, 34 for asthma, and 35 for type 2 diabetes (table 1). These criteria only included aspects of care which the panels judged were necessary to record as well as do, thus enabling medical records to be used to judge quality of care against these criteria. From April 1998 to December 1999 a team from the National Primary Care Research and Development Centre (NPCRDC) undertook a detailed multifactorial quality assessment of a nationally representative sample of 60 randomly selected practices in England. This study included a clinical audit using the previously developed clinical review criteria and a questionnaire survey and interviews with general practitioners and practice nurses to assess their views about the criteria used.

The aim of the clinical audit, postal questionnaire, and interviews was to assess the validity, reliability, and acceptability of the review criteria developed by the expert panels. A review criterion is a systematically developed statement that relates to a measurable aspect of care that is so clearly defined that data were collected for all patients in the sample, prevalence in sample <1%, and inter-rater kappa <0.60. Others have described the desirable characteristics of review criteria.

---

**Table 1** Criteria rated valid by expert panels and used subsequently in quality assessment

<table>
<thead>
<tr>
<th>Condition</th>
<th>(a) Criteria rated valid by expert panels</th>
<th>(b) Criteria for which data were collected for all patients in the sample</th>
<th>(c) Criteria excluded from analyses because prevalence in sample &lt;1%</th>
<th>(d) Criteria excluded from analyses as inter-rater kappa &lt;0.60</th>
<th>(e) Criteria merged with other criteria for analyses</th>
<th>(f) Criteria used in final analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angina</td>
<td>42</td>
<td>26</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Adult asthma</td>
<td>34</td>
<td>26</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>35</td>
<td>31</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>19</td>
</tr>
</tbody>
</table>
METHODS

Sixty practices in England were selected using a three stage process. Three out of the eight English NHS regions were selected to be nationally representative in terms of rurality, socioeconomic deprivation, and geographical population dispersion. From each of these three regions, two health authorities were selected to be representative of their region in terms of rurality and socioeconomic deprivation. Finally, within each of these six authorities, a random sample of 10 practices was selected, stratified in terms of practice size, training status, and socioeconomic deprivation. Where a practice refused to participate, another with similar characteristics was selected.
was chosen at random and invited to participate; 60 out of 75 practices approached (80%) agreed to take part.

**Sampling**

Lists of patients with a confirmed diagnosis of angina, asthma, or type 2 diabetes who were also taking regular medication from a list of the most commonly prescribed drugs for these conditions were generated from computerised records in each practice. Patients had to have been registered with the practice for 2 years to enable sufficient time for “necessary” care to be undertaken. From the lists generated, 20 patients were selected using random numbers, with a further 20 reserves. In some small practices fewer than 20 patients were included as the relevant practice population base was too small. Twenty patients per practice per condition were chosen as it was felt that most practices would have this number of relevant patients for each of the three conditions.

**Data abstraction**

Data were abstracted for up to 20 patients per condition per practice using standardised data abstraction forms (available from the authors). While the time taken to abstract data depended on the complexity of the patient (volume and density of data) and the quality of the medical records (handwriting and whether notes were summarised), data abstraction took on average 20 minutes per patient.

**Timescales**

Many criteria rated valid by the expert panels related to care provided in a specific time period such as the last year. In order to make the data abstraction practical, criteria with no time period attached were restricted to care provided in the previous 5 years, with the exception of two angina criteria as noted in table 2.

**Analysis**

Certain criteria were applicable to all patients, whereas others were applicable only to subgroups—for example, action to be taken if blood pressure exceeded a certain value. Each criterion in table 2 was scored on a 0/1 basis depending on whether necessary care was provided and recorded for individual patients as appropriate.

These binary indicators were analysed using an item response model via the procedure GLLAMM6 using the STATA program.10 12

**Postal questionnaire**

A questionnaire asking respondents to rate the validity of the criteria listed in table 2 was sent to a nurse and doctor in 59 practices; one practice was used as a pilot. Respondents were asked to use an ordinal scale of 1–9 where 9 meant an action was considered necessary and valid for delivering quality of care and 1 meant that it was clearly not necessary and invalid. Respondents were also asked to state on a “yes/no” basis whether they usually recorded the care relating to the criterion in the medical record.

**Interviews**

A member of the research team visited 59 practices, with one practice being used as a pilot, and conducted interviews with a general practitioner and practice nurse (usually the senior partner and nurse) who had completed the postal questionnaire, as well as the practice manager. Interviews took place in September to December 1999 and followed a semi-structured schedule. Interviewees had been sent feedback relating to their practice’s clinical audit results. The second part of the interview discussed respondents’ opinions of the criteria used in the audits, their practice’s results, and possible explanations for low scores.

**RESULTS**

All 60 practices in the main study took part in the clinical audit. Data were collected for 1048, 1133, and 1111 patients with angina, asthma, and type 2 diabetes, respectively. The mean number of patients included in each practice was 18 (range 6–20) for angina, 19 (range 13–20) for asthma, and 18 (range 9–20) for type 2 diabetes.

Doctors and nurses in 56 of the 60 practices completed the questionnaire (response rate 93%) and 55 practices (92%) took part in the interviews. Three practices refused to take part in both parts of the study, and in some practices staff were unavailable for interview or nurse posts were vacant. A total of 161 interviews were conducted (55 with doctors, 51 with nurses, and 55 with practice managers).

**Review criteria rated valid by expert panels that were not be included in the audit**

Not all the review criteria rated valid by the expert panels14 were included in the quality assessment (table 1, column a). Firstly, some criteria were practice level criteria—for example, diabetic register—and these were excluded as they would have been the same for all patients and therefore not discriminated between patients; this excluded two asthma and two diabetes criteria. In addition, screening and diagnostic criteria—for example, family history of angina—were excluded as the audit focused on patients with confirmed diagnoses; this excluded a further 13 angina criteria and one asthma criterion. In addition, some criteria could not be operationalised using an audit abstraction sheet so these items were excluded (three angina, five asthma, two diabetes; box 1).

**Data on quality of care**

Data were subsequently collected for all patients in the relevant samples using 26 criteria for angina, 26 for asthma, and 31 for diabetes (table 1, column b). Data were abstracted from both manual records (including clinic cards and hospital letters) and computerised records.

**Prevalence**

If a criterion applied to fewer than 1% of the relevant condition sample it was not included in any analyses (table 1, column c). This applied to 10 angina criteria, six asthma criteria, and three diabetes criteria. While this cut off point was arbitrary, criteria relevant to fewer than 1% of a condition sample were prone to clustering within practices. These criteria mostly related to combinations of medication—for example, whether verapamil was used in combination with β blockade as second line treatment for patients with angina.

---

Box 1 Examples of review criteria which could not be operationalised

- Whether angina patients had been referred to a cardiologist or for an exercise test at the time of their first prescription at initial diagnosis (initial diagnosis or first prescription may be difficult to identify with certainty).
- Whether a diabetic patient whose feet were “at risk” had been referred to a chiropodist (“at risk” could not be easily operationalised).
- Whether an asthmatic patient on long term maintenance oral steroids had previously had a trial of long acting bronchodilator, high dose inhaled steroids, and one other step 4 British Thoracic Society recommended intervention (this criterion was found to be too complicated and practice prescribing records were often inadequate to extract the data).
Box 2 Examples of items with poor inter-rater reliability

Asthma
• Record of whether maintenance treatment was initiated or increased if the patient was experiencing daily symptoms (such symptoms proved difficult to distinguish from nocturnal and activity limiting symptoms).

Angina
• Record of whether the patient had been referred for an exercise test or to a cardiologist if the patient had not had revascularisation but had had a negative exercise test, was more than minimally symptomatic, and on two drug maintenance treatment (this criterion was found to be too complicated and “more than minimally symptomatic” proved difficult to operationalise).

Type 2 diabetes
• Record of general well being (data abstractors could not agree about what constituted a recorded comment about general well being).

Reliability
Inter-rater reliability, a prerequisite for validity, is the level of agreement between different users of an instrument for the same sample. Two raters abstracted data separately (but on the same day) for 7.5% of all patient records included in each of the three samples across 23 of the practices in the main study. Items with a Cohen kappa coefficient of agreement value <0.60 were excluded from the analyses (table 1, column d). This applied to two angina criteria, three asthma criteria, and five diabetes criteria. Values above 0.6 have good to very good agreement. Examples of items excluded because of poor inter-rater reliability are listed in box 2.

Other omissions
Some criteria—for example, diabetes criteria relating to recording of vibration sense and peripheral pulses—were combined by the research team to create a single criterion (table 1, column e).

Measuring quality of care
The denominators for angina, asthma, and diabetes are 26, 22, and 27, respectively, as columns b and e in table 1 refer to criteria excluded for reasons which are not related to the process of development—for example, practice level criteria or due to merging criteria. Consequently, 54%, 59%, and 70% of the angina, adult asthma and type 2 diabetes criteria rated valid by the expert panels had further evidence of face validity, were feasible to apply, and could be applied reliably. These criteria cover a broad spectrum of care including prevention, evaluation, treatment, and referral (table 2).

Of the 14 angina, 13 asthma and 19 diabetes criteria rated valid, nine, four and 10, respectively, were unconditional criteria relevant to all patients in the relevant sample, whereas five, nine, and nine, respectively, were conditional variables only relevant to a patient dependent upon the answer to another question—for example, action taken if blood pressure exceeded a given value. While some criteria were discarded because of low prevalence, the mean number of criteria relevant to individual patients within the three samples were 10 for angina, five for asthma, and 12 for diabetes. This showed that, on average, only one, one, and two conditional variables, respectively, were relevant to patients in the three samples.

Table 2 shows how frequently each criterion was met for all patients for whom that criterion was relevant, for each condition. Investigations and procedures were more often performed and recorded than advice and prevention. Table 3 shows the variation in scores for individual patients in the three samples.

Validity
The doctors and nurses in the questionnaire survey rated all of the criteria listed in table 2 as valid measures of quality of care (overall median ≥7 without disagreement). All the criteria listed in table 2 were recorded on a consistent basis according to over 80% of all nurses in the questionnaire survey and, with the exception of the seven criteria listed in box 3, by over 80% of all doctors.

Table 3 Necessary criteria not recorded routinely in the medical record by more than 20% of doctors

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean patient score</th>
<th>Minimum patient score</th>
<th>Maximum patient score</th>
<th>Standard deviation of patient scores</th>
<th>Mean number of items relevant to individual patients</th>
<th>Mean number of items met for individual patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angina (n=1048)</td>
<td>60.48</td>
<td>0</td>
<td>100</td>
<td>22.61</td>
<td>10.4</td>
<td>6.4</td>
</tr>
<tr>
<td>Asthma (n=1133)</td>
<td>58.73</td>
<td>0</td>
<td>100</td>
<td>30.2</td>
<td>5.3</td>
<td>3.2</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>71.28</td>
<td>0</td>
<td>100</td>
<td>21.76</td>
<td>12.32</td>
<td>8.79</td>
</tr>
</tbody>
</table>
Secondly, despite agreeing with panellists that all the criteria in table 2 were valid, some respondents disagreed with panel recommendations that all criteria were necessary for all practices. Patient-centred care was seen to be irreconcilable with clinical guideline and protocol based care. For example, while the expert panels explicitly decided that some criteria should be applied to patients of all ages, some interviewees argued that referral for an exercise test or the importance of cholesterol testing were age specific—for example, less meaningful in patients over the age of 80. This reflected differences in perception by the doctors and nurses interviewed and the expert panellists. We have argued elsewhere that quality of care is at its most meaningful when related to individual patients. While the panellists focused upon care for individual patients, in practice there may be a difference between taking part in a consensus method which considers care relevant to an “average” patient with a given condition, and sitting in a consulting room with a patient with complex co-morbidities and personal circumstances.

Other reasons given for poor performance were poor recording by doctors rather than nurses, inadequate or inconsistent information technology, poor data recording templates, insufficient computer training, and poor patient compliance and attendance. However, more frequently, doctors and nurses accepted that poor audit results simply reflected the fact that necessary care had not been taken.

Most staff felt that the audit results for their practice painted an accurate picture of their care for angina, asthma, and diabetes or corroborated other assessments such as those by their local primary care group. However, only a few practices reported at interview that they would change their procedures or protocols as a result of the audit. For example, two practices intended using their results as a baseline as part of bids to become Personal Medical Services pilots and to re-audit their care of angina, asthma, and diabetes. Other staff stated at interview that they had discussed their results within their practice and initiated quality improvement initiatives, including an explicit intention to re-audit care.

DISCUSSION

This study shows that some of the criteria developed previously by expert panels were unoperationalisable, unreliable, too rare to be useful, or too hard to extract reliably. This finding emphasises the fact that quality measures need field testing before they can be used in quality assessment. Nevertheless, the expert panels produced review criteria which were found, after extensive field testing, to be valid and reliable. These findings therefore provide additional evidence that the RAND panel method develops valid and reliable review criteria for assessing clinical quality of care. However, the audit showed clearly that many patients are not receiving necessary care.

The collection of audit data for clinical care represents a cornerstone of many current initiatives such as National Service Frameworks and many primary care groups/trusts are collecting audit data from their practices as part of clinical governance initiatives. These findings have some important implications for the successful implementation of quality improvement in general practice. Firstly, table 2 shows that, while each of the criteria had been rated necessary by both the expert panels and by general practitioners and practice nurses in the practice sample, the care was frequently not provided to patients who needed it. This is perhaps unsurprising as variation in quality of care is endemic in the UK. However, it shows that there is significant room for improvement in the quality of chronic disease management delivered in general practice in the UK.

Secondly, the dominant approach to quality improvement in the UK over the last decade has been audit. However, others have found that only 24% of audits involved a re-audit to see if care had improved, and that only 35% of audit recommendations are implemented. In this study, while most staff felt that their results reflected an accurate picture of their care for angina, asthma, and diabetes, these findings confirm that only a few practices were keen to use the data to improve their care and to re-audit. Those charged with improving quality of care in general practice, particularly primary care groups/trusts, need to motivate practice staff to see the value of auditing and reviewing their care, as many practice staff have a negative attitude towards audit. This will require engaging often suspicious practice staff, as well as cultural and behavioural changes in the attitudes of practice staff.

The fact that the criteria used were acceptable to doctors and nurses in a representative sample of practices in England is, however, important as shared understanding and ownership of ideas enhances successful implementation of change. Thirdly, we found that practices had significantly different levels of computerisation. The increasing availability and comprehensiveness of electronic information systems such as PRODIGY will foster reliable quality assessment, particularly of clinical data. Data need to be reliable, especially if financial incentives or penalties are to result from quality assessments. However, this will require investment to ensure that all practices have comparable data systems.

Limitations

The review criteria were evidence based and they were developed in 1997. The evidence upon which some of the criteria were based is now out of date. For example, diabetes criteria relating to blood pressure control were developed before publication of the UK PDS study. This emphasises the importance of updating review criteria and the evidence/literature reviews upon which they are based.

There is some evidence that the quality of record keeping is positively correlated with increased quality of care. However, there has been concern about the validity and reliability of using medical records to assess quality of care; in particular, that data abstraction from records underestimates quality of care because records are not sensitive enough to measure all that goes on in a consultation, especially preventive or counselling/advice activities. This limitation was often emphasised by the general practitioners and nurses during interviews. Certainly, while most of the criteria used in this study focused on clinical care, table 2 shows that criteria relating to preventive care and symptoms were less frequently met than those pertaining to investigations/procedures. It is not possible to state how frequently care was given but not recorded or simply not given at all; audit does not distinguish between the two. It is important to emphasise that poor audit results can either reflect poor care or poor recording—a fact accepted by general practitioners and practice nurses. The difference between what the expert panels recommended as
valid review criteria and the views of what doctors and nurses in practices think should always be recorded is an important outcome of the study.

In conclusion, the criteria reported in table 2 provide a workable group of review criteria that could be used by primary care organisations and general practitioners for assessment of the quality of care they deliver to patients with angina, adult asthma, or type 2 diabetes.

ACKNOWLEDGEMENTS

The authors thank the staff of all 60 practices and six health authorities which took part in the study and acknowledge Cath Burns, Dianne Oliver, Nicola Mead and Emma Ruff for their contribution to this project.

References

2. Roland MO, Holdin J, Campbell S. Quality assessment for general practice: supporting clinical governance in primary care groups. Manchester: National Primary Care Research and Development Centre, University of Manchester, Manchester M13 9PL, UK
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
S M Campbell, M Hann, J Hacker, A Durie, A Thapar and M O Roland

Qual Saf Health Care 2002 11: 125-130
doi: 10.1136/qhc.11.2.125