Assessing quality of care from hospital case notes: comparison of reliability of two methods

A Hutchinson,[1] J E Coster,[1] K L Cooper,[1] A McIntosh,[1] S J Walters,[2] P A Bath,[3] M Pearson,[4] K Rantell,[2] M J Campbell,[2] J Nicholl,[2] P Irwin[4]

► Supplementary table E1 is published online only. To view these files please visit the journal online (http://qshc.bmj.com).

[1]Section of Public Health, ScHARR, University of Sheffield, Sheffield, UK [2]Section of Health Services Research, ScHARR, University of Sheffield, Sheffield, UK [3]Department of Information Studies, University of Sheffield, Sheffield, UK [4]Clinical Effectiveness and Evaluation Unit, Royal College of Physicians, London, UK

Correspondence to
Allen Hutchinson, Section of Public Health, ScHARR, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK; allen.hutchinson@sheffield.ac.uk

Accepted 20 January 2009

ABSTRACT

Objectives To determine which of the two methods of case note review provide the most useful and reliable information for reviewing quality of care.

Design Retrospective, multiple reviews of 692 case notes were undertaken using both holistic (implicit) and criterion-based (explicit) review methods. Quality measures were evidence-based review criteria and a quality of care rating scale.

Setting Nine randomly selected acute hospitals in England.

Participants Sixteen doctors, 11 specialist nurses and three clinically trained audit staff, and eight non-clinical audit staff.

Analysis Methods Intrarater consistency, inter-rater reliability between pairs of staff using intraclass correlation coefficients (ICCs), completeness of criterion data capture and between-staff group comparison.

Results A total of 1473 holistic reviews and 1389 criterion-based reviews were undertaken. When the three same staff types reviewed the same record, holistic scale score inter-rater reliability was moderate within each group (ICC 0.46 to 0.52). Inter-rater reliability for criterion-based scores was moderate to good (ICC 0.61 to 0.88). Comparison of holistic review score and criterion-based score of case notes reviewed by doctors and by non-clinical audit staff showed a reasonable level of agreement between the two methods.

Conclusions Using a holistic approach to review case notes, same staff groups can achieve reasonable repeatability within their professional groups. When the same clinical record was reviewed twice by the doctors, and by the non-clinical audit staff, using both holistic and criterion-based methods, there are close similarities between the quality of care scores generated by the two methods. When using retrospective review of case notes to examine quality of care, a clear view is required of the purpose and the expected outputs of the project.

Quality of care is assessed from clinical records using two principal approaches: holistic (implicit) and criterion-based (explicit) review. Both approaches have strengths and weaknesses, whether they are used for performance monitoring, assessment or research.

Clinical staff are accustomed to reviewing patient records to judge the quality of care. This holistic approach uses professional judgement and requires no prior assumptions about the individual case, can be applied to any condition, can extend to examining any aspect of care and may be relatively quick. However, the standards against which quality is judged holistically are implicit, being dependent on the reviewer's personal knowledge and perspective, and therefore subjective.

Semistructured holistic review methods have therefore been developed to determine standards of hospital, outpatient and nursing care.[1–3] These methods ask specific questions about phases and aspects of care and may use scales to rate quality.

Nevertheless, despite attempts to reduce levels of subjectivity in holistic review by providing extensive training for physician reviewers, concerns remain about review methods based principally on professional judgement. There are concerns about inter-rater reliability,[4] choice of methods of assessing reliability,[5] consistency,[6] bias towards harshness or leniency,[7] hindsight bias[8] and reviewer idiosyncrasy.[9] Moreover, lower levels of inter-rater reliability have been found for holistic review than for criterion-based review.[9] Criterion-based review has therefore been proposed as a more effective means of assessing quality.[10 11]

Criterion-based review allows comparison of care against explicit standards (eg, from national clinical guidelines), where unambiguous questions are defined to construct variables with good reproducibility, for retrieval from case notes. Clinical audit in the UK has adopted these objective, criterion-based, approaches,[12–14] using explicit standards, independent of profession. These have been used to identify substantial variations in organisation and clinical care between hospitals.[12]

However, criterion-based review has been criticised as being insensitive[15] and may not identify unexpected factors influencing outcomes of care.[16] Mixed methods are an alternative,[17 18] whereby nurses use criterion-based review to identify a subset of problematic cases for subsequent holistic review by doctors; however, prior selection may lead to hindsight bias among the physician reviewers who may judge selected cases more harshly.[8 11] Moreover, nurses and doctors may use different information to judge care quality (and may make different judgements about an individual case).[18]

It is therefore not clear which review method provides more reliable and useful information or how relatively reliable and reproducible are the different methods when carried by different healthcare professionals. Our study compares the results of three different professional groupings when evaluating quality of care from the same set of case notes, using both holistic review using quality of care rating scales and criterion-based review.

METHODS

Setting and reviewer professional background

Data were collected from nine acute hospitals in England, selected randomly from 136 that met high patient-throughput criteria for the two study conditions. In each hospital, staff were recruited to undertake reviews of cases of an admission for an exacerbation of either chronic obstructive pulmonary disease (COPD) or heart failure. Three staff types were recruited: 16 doctors in specialist training, 14 other clinical staff (11 of whom were nurses specialising in the review condition) and 8 non-clinical audit staff.

Training and data collection

Reviewers received a 1-day joint training session on holistic and criterion-based review methods. Clinical scenarios were used and reviewers were provided with copies of national clinical guidelines for COPD and heart failure care.[19 20] Data collection software was demonstrated. Reviewers evaluated the records within their own hospital, similar to local clinical audit, and no patient-identifiable data were used in the analysis.

Review methods

Different combinations of reviewers from the three staff types were used at each hospital to compare their effectiveness in carrying out holistic and criterion-based case note reviews. In each hospital, case notes of 50 consecutive admissions of COPD or heart failure were sought and reviewed by staff type combinations of one to four staff (figure 1).

Each reviewer evaluated care on the same case notes using both review methods, holistic and then criterion-based review, holistic being used first to reduce potential hindsight bias[8 11] caused by finding a low criterion-based score first. Using their own implicit standards, reviewers rated the reported quality of care provided to each patient for three structured phases of care (admission/investigations, initial management and predischarge care). Each phase was rated on a 1 to 6 scale (1=unsatisfactory, 6=very best care). Overall quality of care for each review was rated on a 1 to 10 scale (1=unsatisfactory, 10=very best care).

Reviewers then undertook a criterion-based review on the same case notes. Criteria development used established methods for constructing explicit evidence-based review criteria[4 12 13 21] (COPD, n=37; heart failure, n=33) derived from national clinical guidelines recommendations and expert opinion.[19 20]

Criterion-based data were used to assess each reviewer's effectiveness at abstracting data from clinical records and completing the data collection form; an "effectiveness of reviewer" score was calculated and converted to a percentage for each record review (one point per data field completed by the reviewer; one point subtracted per data field left blank). Quality of care scores were calculated for each record, comprising the percentage of the criteria identified by the reviewer as having been met.

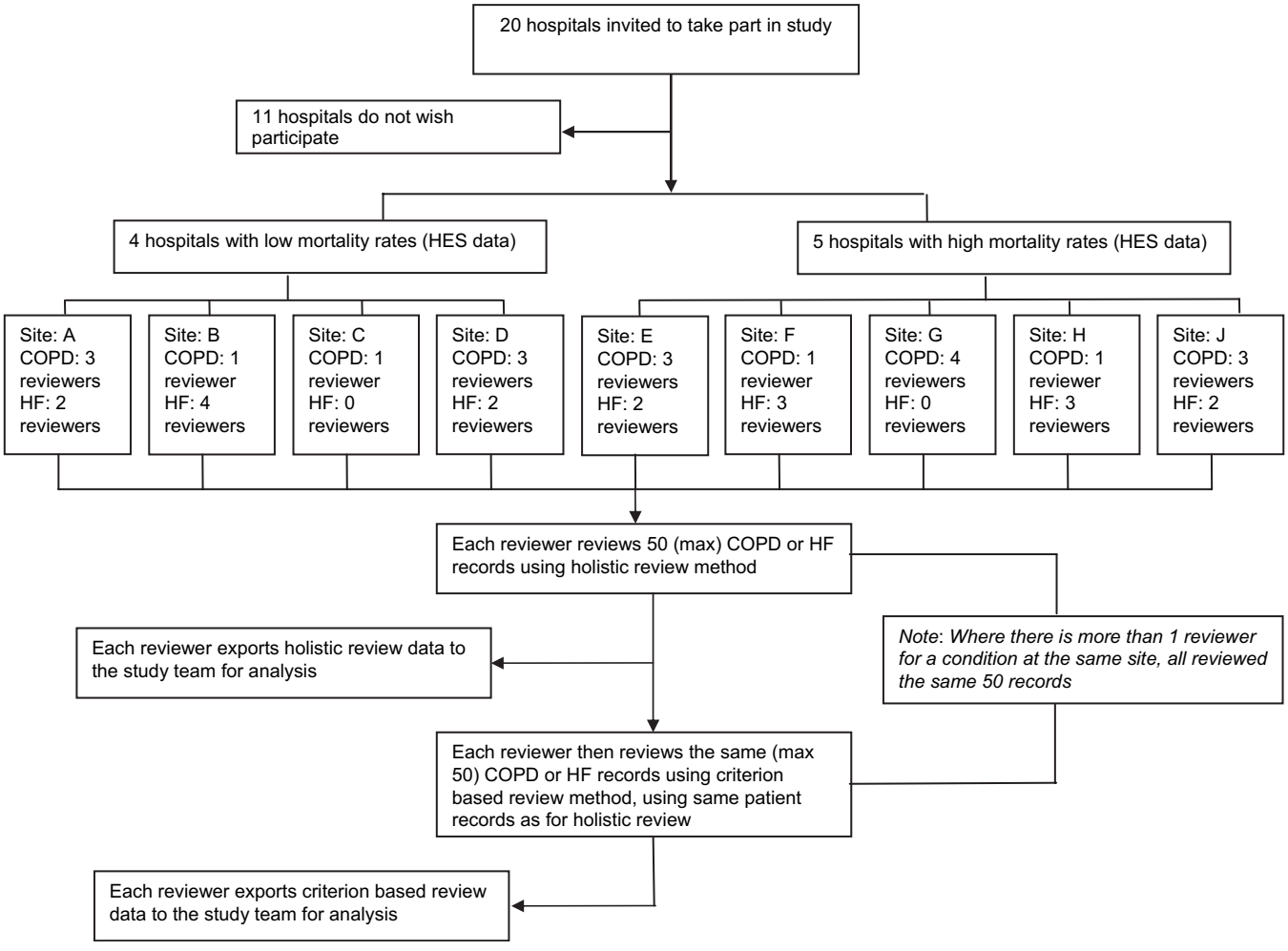

Figure 1 Overview of selection and review process.

Analysis methods

Holistic review

Intrarater and inter-rater reliability was calculated for holistic quality of care scores. Robust standard errors (STATA V.9, College Station, Texas, USA)[22] were used to allow for clustering of scores around each reviewer when calculating confidence intervals and p values for the mean overall scores by reviewer type.

Intrarater consistency for each review was assessed by calculating Pearson's correlation coefficient between the mean rating of the three phases of care and the rating of overall care.

To assess inter-rater reliability between ratings of the same record by different reviewers, raw ratings were converted to ranks to adjust for variation in the range of scores used by different reviewers. Intraclass correlation coefficients (ICCs) were calculated on these ranks.[4 23] The ICC is the correlation between two measurements or quality of care ratings in the same patient, using randomly chosen reviewers.

Criterion-based review

Mean criterion-based quality of care scores were compared across the three staff types using a one-way analysis of variance, taking account of clustering by staff type.

Inter-rater reliability for overall quality of care scores by pairs/triplets of staff reviewing the same records was estimated using ICCs. Pooled ICC estimates from the different combinations of reviewers used a weighting that was inversely proportional to the variance of the estimate.[23]

Inter-rater reliability results for the two review methods were compared.

RESULTS

Across nine acute hospitals, 38 reviewers undertook 1473 holistic reviews and 1389 criterion-based reviews (total=692 case notes). The number of case notes reviewed by each individual ranged from 9 to 50 (see electronic table E1). This variation was due to the effect of job rotations, workload pressures and difficulties in obtaining clinical records.

Intrarater consistency in holistic reviews

For all three staff types (table 1), there were statistically significant correlations ($r>0.71$, $p<0.001$) between the mean scale score ratings that reviewers assigned to the individual phases of care and their rating of the overall quality of care for the same set of case notes, indicating a fair to good level of intrarater consistency in rating the quality of care using holistic review scale scores.

Criterion-based reviewer effectiveness

Effectiveness in capturing criterion-based data was high and similar across all staff types (table 2), with mean scores approximately 95% (approximately 1.5 data items missing per review).

Table 2 Criterion-based reviewer effectiveness scores

Review staff type (no. of review staff)	Number of reviews	Mean score %, SD (95% CI)	Range
Doctor (16)	477	94.9, 4.8 (93.2 to 96.5)	74.2 to 100.0
Nurse/other clinical (14)	443	95.2, 4.1 (93.5 to 97.0)	67.7 to 100.0
Non-clinical audit (9)	289	94.7, 5.0 (93.2 to 96.5)	61.3 to 100.0
Total (39)	1209	95.0, 4.6 (94.0 to 95.9)	61.3 to 100.0

Analysis excludes patients who died.
(95% CI)* are adjusted for clustering by reviewer.

Inter-rater reliability for holistic review

Holistic review reliability between scale score ratings of the same record by pairs of reviewers was moderate within all three staff types, although it varied between reviewer pairs and was sometimes very poor (table 3).

The overall weighted mean ICC was moderate across all three reviewer types, with overlapping 95% confidence intervals (CIs) indicating no significant differences between staff types.

Comparisons between professional groups

Where reviewers from different staff types used holistic scale score methods to review the same record, inter-rater reliability was assessed between staff groups for all phases of care and overall care (table 4). For the holistic phase of care findings within staff groups, there was generally modest to fair agreement within pairs, particularly among doctors, although the range is large even among them (eg, initial management results). However, where staff from different groups reviewed the same record, agreement between the different professional groups on their assessment of the quality of care was poor to non-existent.

Analysis of variance between the holistic overall scale ratings of the three staff types show that the nurse/other clinical group scores were significantly lower than the doctor ($p<0.001$) and non-clinical audit groups ($p<0.001$). The comparison of the latter two groups showed no significant differences ($p=0.352$).

Inter-rater reliability for criterion-based review

Inter-rater reliability between criterion-based scores (ie, the percentage of criteria recorded as being met) for the same record by different reviewers ranged from moderate to good within all staff types. Doctors showed a significantly higher level of reliability (table 5).

Comparison of holistic and criterion-based methods

Inter-rater reliability results for the two review methods were compared. In addition, an estimate of the within-staff-type consistency across both review methods was calculated using p value for difference between the overall holistic quality of care ratings (percentage) and the percentage of criteria recorded as being met.

Table 6 shows that the mean overall "quality of care" scores across the 692 patient records were similar for holistic and

Table 1 Intrarater consistency between holistic scale score ratings for phases of care and for overall care

Review staff type (number of review staff)	Number of reviews*	Mean overall rating of quality of care† (SD)	Mean rating of phase quality of care (based on the mean score across three phases of care)‡	Pearson correlation between mean rating across three phases of care and overall rating
Doctors (16)	593	7.8 (1.8)	4.7 (0.8)	0.77
Nurses/other clinical (14)	529	7.0 (2.0)	4.4 (1.0)	0.81
Non-clinical audit (9)	296	7.9 (1.3)	4.6 (0.8)	0.71

*Numbers of reviews used in tables 1 and 2 differ slightly because of small amounts of missing data because some patients died during the admission and some phases of care were therefore not rated.
†Overall quality of care was rated on a 1 (unsatisfactory) to 10 (very best care) scale.
‡Quality of care in each of the three phases: (admission/investigations, initial management and predischarge) was rated on a 1 (unsatisfactory) to 6 (very best care) scale.

Table 3 Inter-rater reliability between holistic overall ratings of the same record by paired reviewers of different staff types

Reviewer pairs	Condition	Site*	No. of paired reviews	ICC between ranked scores (95% CI)	Weighted mean ICC† (95% CI)
Doctor vs doctor	Heart failure	B‡	49	0.67 (0.54 to 0.79)	0.52 (0.41 to 0.62)
	COPD	G	48	0.33 (0.05 to 0.56)	
	Heart failure	F	18	−0.03 (−0.48 to 0.43)	
	Heart failure	E§	12	−0.44 (−0.80 to 0.15)	
Nurse/clinical vs nurse/clinical	Heart failure	D	21	0.74 (0.47 to 0.89)	0.46 (0.34 to 0.59)
	COPD	D	49	0.37 (0.10 to 0.58)	
	COPD	J	26	0.27 (−0.12 to 0.59)	
	Heart failure	H	48	0.22 (−0.07 to 0.47)	
Non-clinical audit staff vs non-clinical audit staff	COPD	A	48	0.47 (0.22 to 0.66)	0.47 (0.22 to 0.66)

COPD, chronic obstructive pulmonary disease; ICC, intraclass correlation coefficient.
*Only sites with more than one reviewer of the same staff type are included in this table.
†Mean ICC per staff type, weighted by inverse variances to account for differing numbers of paired reviews.
‡A single ICC was calculated for the three doctors at site B.
§The doctors at site E were non-specialist doctors.

criterion-based methods and for all three staff types (70% to 79%, where 100%=excellent care).

Estimation of the level of quality of care score agreement between the two methods for an individual record, using p value for difference, shows that there was no significant difference between the holistic and criterion-based assessments when undertaken by the doctors (mean difference −1.9, p value for difference 0.406) and by the non-clinical audit staff (mean difference 3.1, p value for difference 0.223).

A non-significant p value for difference indicates that there is some association between the scores derived from the two review methods. These results suggest that for the doctors and the non-clinical audit staff the two methods are giving, on average, a somewhat similar result. The pooled results for all staff showed a small mean difference (−2.6) that bordered on statistical significance, possibly influenced by the highly significant results from the nurse/other clinical group (39% of all of the reviews).

DISCUSSION

Retrospective assessment of the quality and safety of care can be performed from the clinical record using holistic or criterion-based review methods: both have methodological constraints. Studies mostly compare different professional groups using different methods. Thus, Weingart *et al*[18] compared explicit (criterion-based) review undertaken by nurses with implicit review of the same record undertaken by physicians, and found that "nurse and physician reviewers often came to substantially different conclusions". This is the first UK study to contrast the two methods of review systematically and also across three different professional groups.

We investigated the level of agreement between healthcare professionals, from different backgrounds, when they review the same record. This agreement, or reliability, relates to the repeatability of the results from the review—whether a different reviewer would come to the same conclusion about the quality of care from the same data source, using the same method. This is clearly a practical question for those reviewing quality of care in clinical audit or performance review.

Reviewers undertaking holistic review, using scale scores, were relatively consistent in the scores allocated to care quality across the individual phases of care and overall for the entire episode of care. All three staff groups had moderate within-group inter-rater reliability, ranging from 0.46 (95% CI 0.34 to 0.59) to 0.52 (95% CI 0.41 to 0.62), with the doctor reviewers faring best. These were rather higher values than the average found in a systematic review by Lilford *et al*,[5] in which implicit structured case note review studies concerned with causality and process of care had mean κ values <0.4 (causality; κ 0.39 (SD 0.19), process; κ 0.35 (SD 0.19)). Our study results are also somewhat similar to

Table 4 Within-staff-type ICC and between-staff-type group ICC comparisons of holistic scale score reliability for phases of care and overall score

Reviewer pairs	No. of reviewer pairs (or triplets)	No. of case notes		Weighted mean ICC* between ranked scores: Admission/ investigations and examination phase	Initial management phase	Predischarge phase	Overall care
Within-staff-type ICC results							
Doctor vs doctor	4	127	Weighted mean	0.58	0.70	0.46	0.52
			95% CI	0.48 to 0.68	0.63 to 0.78	0.34 to 0.59	0.41 to 0.62
Nurse/clinical vs nurse/clinical	4	144	Weighted mean	0.50	0.22	0.43	0.46
			95% CI	0.38 to 0.62	0.07 to 0.37	0.30 to 0.55	0.34 to 0.59
Non-clinical audit staff vs non-clinical audit staff	2	87	Weighted mean	0.35	0.10	0.39	0.47
			95% CI	0.16 to 0.54	−0.10 to 0.30	0.21 to 0.57	0.22 to 0.66
Between staff type comparisons of ICC							
Doctor vs nurse/clinical	5	179	Weighted mean	0.23	0.25	0.29	0.43
			95% CI	0.09 to 0.37	0.12 to 0.39	0.16 to 0.43	0.31 to 0.54
Doctor vs non-clinical audit staff	6	188	Weighted mean	−0.01	0.03	0.25	0.24
			95% CI	−0.15 to 0.12	−0.11 to 0.16	0.12 to 0.38	0.12 to 0.37
Nurse/clinical vs non-clinical audit staff	1	34	Weighted mean	−0.12	0.19	0.47	0.43
			95% CI	−0.44 to 0.23	−0.15 to 0.49	0.17 to 0.70	0.11 to 0.67

ICC, intraclass correlation.
*Weighted mean ICC: estimates from the different combinations of reviewers were pooled using a weighting that was inversely proportional to the variance of the estimate.

Table 5 Inter-rater reliability between criterion-based scores (proportion of criteria stated as being met) for the same record by different reviewers

Reviewer pairs	Condition	Sit*	No. of paired reviews	ICC between scores (95% CI)	Weighted mean ICC† (95% CI)
Doctor vs doctor	Heart failure	F	14	0.96 (0.87 to 0.99)	0.88 (0.83 to 0.93)
	COPD	G	50	0.65 (0.46 to 0.79)	
	Heart failure	B	46	0.65 (0.50 to 0.77)	
	Heart failure	E‡	12	0.64 (0.13 to 0.88)	
Nurse/clinical vs nurse/clinical	COPD	J	25	0.86 (0.71 to 0.94)	0.74 (0.66 to 0.82)
	COPD	D	48	0.70 (0.52 to 0.82)	
	Heart failure	D	21	0.69 (0.38 to 0.86)	
	Heart failure	H	50	0.27 (0.00 to 0.51)	
Non-clinical audit staff vs non-clinical audit staff	COPD	E	40	0.69 (0.49 to 0.82)	0.61 (0.47 to 0.76)
	COPD	A	29	0.33 (−0.04 to 0.61)	

COPD, chronic obstructive pulmonary disease; ICC, intraclass correlation.
*Only sites with more than one reviewer are included in reliability analysis; therefore, some sites do not appear on this table.
†Mean ICC per staff type, weighted by inverse variances to account for differing numbers of paired reviews. A single ICC was calculated for the three doctors at site B and this was combined with the other doctor pairs in the weighted mean ICC.
‡Non-specialist doctors.

those of Hofer *et al*[4] who used ICCs to examine repeatability and found a reliability of 0.46 for a structured holistic review of diabetes and heart failure case notes by physician reviewers (although only 0.26 for case notes of patients with COPD). By comparison, a recent holistic assessment of patients dying in UK hospitals achieved a κ score of 0.39 on the key indicator of quality of medical care.[24]

Criterion-based review demonstrated that all reviewers could identify relevant data (the effectiveness of reviewer scores were around 95%). There were moderate (0.61 for non-clinical audit staff) to quite high levels of inter-rater reliability (clinical staff 0.74, doctors 0.88)—similar to those found in large UK national clinical audit programmes of stroke[25 26] and continence,[27] and reflecting the trend to higher values for explicit reviews found in other studies.[5] Our study confirms the findings of the UK stroke care audit,[25 26] that criterion-based record review can be undertaken by staff from different backgrounds.

Case note review can only consider what has been recorded, and incomplete records do not mean that an event did not occur. If a practitioner considered something too trivial to record, then it is doubtful that any consequential actions would have occurred. However, some significant events will remain unrecorded and thus unreviewed. Direct observation of care delivery overcomes the problem of missing information, and is an alternative approach,[17] although too expensive as a standard procedure. Hindsight bias in case note review is an acknowledged challenge.[28] We tried to minimise any effect by undertaking holistic review before criterion review.

The overall results of care quality assessment were similar with both methods from our review and all rated care quality reasonably highly (between 70% and 79%, where 100% represents excellent care). But the weak inter-group reliability for holistic scores has implications when choosing how to evaluate the care quality from case notes. Performing as a screening tool, criterion-based review produces sufficient information to judge the overall quality of care, provided that appropriate review criteria are chosen. A structured form of holistic review also gives a reliable picture of the quality of care in the right hands, yet can also pick up extra nuances of quality variation.

Our medical reviewers were relatively inexperienced but with audit training were able to use both criterion-based and holistic review effectively. It would be interesting to explore whether senior clinicians' greater clinical experience would produce different holistic assessments. We hypothesise that it would useful to explore further the expertise of specialist nurses in holistic review because they have particular skills in helping patients with adherence to care pathways.

So which method of review would be best used for clinical audit and performance review, and by which professional groups? All three professional groups performed well when using criterion-based review, so the decision on who should undertake reviews depends mainly on cost and availability of staff.

On the other hand, the decision on who should undertake structured holistic review is more complex. The method can deliver more than just the sum of the results of collecting a set of review criteria. Although all groups can use the method of holistic scale scoring, our data suggest that, for the more technical phases of care, the three groups interpreted the same records differently despite considerable training in the review method. To some extent this probably reflects their background knowledge of clinical care delivery. It is unrealistic to expect non-clinical audit staff to fully appreciate the details of the medical care, let alone judge when care has deviated from best practice.

Table 6 Mean ratings/scores of overall quality of care: comparison of two review methods

Staff type	No. of holistic and criterion-based reviews (and review staff)*	Holistic mean rating of overall quality of care† (95% CI)	Criterion-based review mean score as a percentage of total criteria‡ (95% CI)	Mean difference (95% CI)	p Value for difference
Doctor	462 (16)	76.8 (72.2 to 81.4)	78.7 (77.1 to 80.4)	−1.9 (−6.7 to 2.9)	0.406
Nurse/other clinical	428 (14)	71.2 (66.4 to 76.0)	77.5 (75.0 to 80.1)	−6.3 (−10.5 to −2.2)	0.005
Non-clinical audit	219 (8)	78.5 (74.7 to 82.3)	75.4 (71.1 to 79.7)	3.1 (−2.4 to 8.5)	0.223
All staff	1109 (38)	75.0 (72.3 to 77.6)	77.6 (76.2 to 79.0)	−2.6 (−5.4 to 0.1)	0.057

All CIs and p values are adjusted for clustering by staff type.
*Numbers of reviews used in tables 1 and 2 differ slightly because of small amounts of missing data.
†Reviewers rated the overall quality of care on a 10-point scale from 1 (unsatisfactory) to 10 (very best care). This was converted to a percentage for comparison with criterion-based review data.
‡Scores are shown as percentages out of 32 criteria (where patient is a current or ex-smoker) or out of 31 criteria (where patient is a non-smoker).

Although nurses are much closer to the medical care process, the limited agreement between the doctors and the nurses may reflect different internal professional standards for assessing quality and safety of care. Weingart *et al*[18] conjectured that nurses and doctors reviewed in different ways, that nurses sought data on the routines of care while doctors looked for a wider picture and that neither group considered both dimensions. Analysis of textual commentary on quality of care available from each holistic review will throw further light on this question.

CONCLUSIONS

There is modest agreement between the holistic and criterion-based quality assessment scores of the same record by the same reviewer. However, for holistic review, different staff groups are implicitly using different care standards in their assessment of quality. Large-scale criterion-based audits, such as those promoted by the English Healthcare Commission,[29] may miss the richer information provided by holistic review. A mixed holistic and criterion-based approach may be a solution[5] and has been subsequently used in this study to investigate the relationship between care process and outcome.

Acknowledgements Karen Beck has provided administrative support throughout the project and her contribution has been exceptionally helpful.
The research team would like to thank all of those NHS staff who so generously contributed to the successful completion of this study. We also wish to thank staff at the Royal College of Physicians Clinical Effectiveness and Evaluation Unit for their help in selecting research sites and methods development.

Funding DoH Methodology Research Programme c/o Department of Epidemiology and Public Health, University of Birmingham, Edgbaston, Birmingham, UK. The project was funded by the Department of Health for England Methodology Research Programme.

Competing interests None.

Ethics approval An opinion was sought from the Trent Multiple Research Ethics Committee and the project was deemed not to require formal ethical review because no identifiable individual patient data were seen by the research team.

Contributors AH developed and directed the study and acted as lead author for this paper. JED project managed the study, collected and analysed the data. KLC assisted with project management, undertook recruitment and data analysis. AMC acted as senior methodologist and lead qualitative researcher. SJW lead the statistical analysis and contributed to methods development. PAB contributed to methods development and analysis. MP acted as senior clinician, contributing to recruitment, methods development, analysis and writing. KR undertook statistical analysis and contributed to methods development. MJC acted as senior statistical adviser. JPN contributed to the development of the study and to the analytic framework. PI contributed to methods development and to recruitment. All authors contributed to the writing of this paper.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Agency for Health Care Policy and Research.** Using clinical practice guidelines to evaluate quality of care. **Vol 2**. Rockville (MD): US Department of Health and Human Services, Public Health Service, 1995.
2. **Rubenstein LR,** Kahn KL, Harris ER, *et al*. Structured implicit review of the medical record: a method for measuring the quality of inhospital medical care and a summary of quality changes following implementation of the Medicare prospective payments system. Santa Monica: RAND, 1991
3. **Pearson M,** Lee JL, Chang BL, *et al*. Structured implicit review: a new method for monitoring nursing care quality. *Med Care* 2000;**38**:1074—91.
4. **Hofer TP,** Asch SM, Hayward RA, *et al*. Profiling quality of care: is there a role for peer review? *BMC Health Serv Res* 2004;**4**:9. http://www.biomedcentral.com/1472-6963/4/9 (accessed 28 Nov 2007).
5. **Lilford R,** Edwards A, Girling A, *et al*. Inter-rater reliability of case-note audit: a systematic review. *J Health Serv Res Policy* 2007;**12**:173—80.
6. **Hulka BS,** Romm FJ, Parkerson GR, *et al*. Peer review in ambulatory care: use of explicit criteria and implicit judgements. *Med Care* 1979;**17**:1—73.
7. **Hayward RA,** McMahon LF, Bernard AM. Evaluating the care of general medicine inpatients: how good is implicit review? *Ann Intern Med* 1993;**118**:550—6.
8. **Fischoff B.** Hindsight not equal to foresight: the effect of outcome knowledge on judgement under uncertainty. *J Experimental Psychol* 1975;**1**:288—99.
9. **Ashton C,** Kuykendall D, Johnson ML, *et al*. An empirical assessment of the validity of explicit and implicit process of care criteria for quality assessment. *Med Care* 1999;**37**:798—808.
10. **Localio RA,** Weaver SL, Landis R, *et al*. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. *Ann Intern Med* 1996;**125**:457—64.
11. **Hayward RA,** Hofer TPE. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA* 2001;**286**:415—20.
12. **Rudd AG,** Lowe D, Irwin P, *et al*. Intercollegiate Stroke Working Party. National stroke audit: a tool for change? *Qual Health Care* 2001;**10**:141—51.
13. **Hutchinson A,** McIntosh A, Anderson J, *et al*. Developing primary care review criteria from evidenced-based guidelines: coronary heart disease as a model. *BJGP* 2003;**53**:691—6.
14. **The North of England Study of Standards and Performance in General Practice.** Medical audit in general practice. I: Effects on doctors' clinical behaviour for common childhood conditions. *BMJ* 1992;**304**:1480—4.
15. **Camacho LA,** Rubin HR. Assessment of the validity and reliability of three systems of medical record screening for quality of care assessment. *Med Care* 1998;**36**:748—51.
16. **Mohammed MA,** Mant J, Bentham L, *et al*. Process and mortality of stroke patients with and without do not resuscitate order in the West Midlands, UK. *Int J Qual Healthcare* 2006;**18**:102—6.
17. **Thomas EJ,** Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. *Ann Intern Med* 2002;**136**:812—16.
18. **Weingart SN,** Davis RB, Palmer RH, *et al*. Discrepancies between explicit and implicit review: physician and nurse assessments of complication and quality. *Health Serv Res* 2002;**32**:483—98.
19. **National Institute for Clinical Excellence.** Chronic obstructive pulmonary disease. Management of chronic obstructive pulmonary disease in adults in primary and secondary care. London: National Institute for Clinical Excellence, 2004.
20. **National Institute for Clinical Excellence.** Chronic heart failure: management of chronic heart failure in adults in primary and secondary care. Clinical Guideline 5. London: National Institute for Clinical Excellence, 2003
21. **Hadorn DC,** Baker DW, Kamberg CJ, *et al*. Practice guidelines. Phase II of the AHCPR-sponsored heart failure guideline: translating practice recommendations into review criteria. *J Qual Improv* 1996;**22**:265—76.
22. **StataCorp.** Stata statistical software: release 9. College Station (TX): StataCorp LP, 2005.
23. **Fleiss JL.** Statistical methods for rates and proportions. 2nd edn. New York: Wiley, 1981.
24. **Wardle TD,** Burnham R, Greig E, *et al*. A confidential study of deaths after emergency medical admission: issues relating to quality of care. *Clin Med* 2003;**3**:425—34.
25. **Gompertz PH,** Irwin P, Morris R, *et al*. Reliability and validity of the Intercollegiate Stroke Audit Package. *J Eval Clin Prac* 2001;**7**:1—11.
26. **Gompertz P,** Dennis M, Hopkins A, *et al*. Development and reliability of the stroke audit form. UK Stroke Audit Group. *Age Aging* 1994;**23**:378—83.
27. **Potter J,** Peel P, Mian S, *et al*. National audit of continence care for older people: management of faecal incontinence. *Age Ageing* 2007;**36**:268—73.
28. **Lilford RJ,** Mohammed MA, Brauholtz D, *et al*. The measurement of active errors: methodology issues. *QSHC* 2003;**12**:ii8—ii12.
29. http://www.healthcarecommission.org.uk/national clinical audit/ (accessed 28 Nov 2008).