Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires

Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires

J L Campbell, S H Richards, A Dickens, M Greco, A Narayanan and S Brearley

Qual. Saf. Health Care 2008;17;187-193
doi:10.1136/qshc.2007.024679

Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires

J L Campbell,[1] S H Richards,[1] A Dickens,[1] M Greco,[2] A Narayanan,[2] S Brearley[3]

► Additional data are published online only at http://qshc.bmj.com/content/vol17/issue3

[1] Primary Care Research Group, Peninsula Medical School, Exeter, UK; [2] CFEP-UK Surveys, Exeter, UK; [3] UK General Medical Council, London, UK

Correspondence to:
J L Campbell, Primary Care Research Group, Peninsula Medical School, Smeall Building, St Luke's Campus, Magdalen Road, Exeter EX1 2LU, UK; john.campbell@pms.ac.uk

Accepted 2 March 2008

ABSTRACT

Objective: To investigate the utility of the GMC patient and colleague questionnaires in assessing the professional performance of a large sample of UK doctors.

Design: Cross-sectional questionnaire surveys.

Setting: Range of UK clinical practice settings.

Participants: 541 doctors gave preliminary agreement to take part in the study. Responses were received from 13 754 patients attending one of 380 participant doctors, and from 4269 colleagues of 309 participant doctors.

Main outcome measures: Questionnaire performance and standardised scores for each doctor derived from patient and colleague responses.

Results: Participant doctors were similar to non-participants in respect of age and gender. The patient and colleague questionnaires were acceptable to participants as evidenced by low levels of missing data. One patient questionnaire item seemed to cause confusion for respondents and requires rewording. Both patient and colleague responses were highly skewed towards favourable impressions of doctor performance, with high internal consistency. To achieve acceptable levels of reliability, a minimum of 8 colleague questionnaires and 22 patient questionnaires are required. G coefficients for both questionnaires were comparable with internationally recognised survey instruments of broadly similar intent. Patient and colleague assessments provided complementary perspectives of doctors' performance. Older doctors had lower patient-derived and colleague-derived scores than younger doctors. Doctors from a mental health trust and doctors providing care in a variety of non-NHS settings had lower patient scores compared with doctors providing care in acute or primary care trust settings.

Conclusions: The GMC patient and colleague questionnaires offer a reliable basis for the assessment of professionalism among UK doctors. If used in the revalidation of doctors' registration, they would be capable of discriminating a range of professional performance among doctors, and potentially identifying a minority whose practice should to subjected to further scrutiny.

Recent years have seen increasing interest across many healthcare systems in the assessment of the professional performance of doctors with a view to establishing processes for their professional revalidation or re-licensing.[1] Despite this, only limited published evidence is available regarding the reliability and effectiveness of such processes.

In 1998 the General Medical Council (GMC), which registers and regulates doctors practising in the UK, determined that "all doctors should be prepared to demonstrate at regular intervals that they remain up to date and fit to practise",[2] and shortly afterwards proposed that participation in such a process should become a condition of continued registration. Attempts to translate these principles in to a workable method for the revalidation of doctors' have been hampered by the lack of reliable methods of assessing doctors' competence and performance. Measures of outcome have proved difficult to interpret because of variations in case mix and the problem of attributability.[3] Tests of knowledge may not be congruent with individual doctors' fields of practice, and may correlate poorly with both competence and performance.[4]

In an attempt to overcome such difficulties, many individuals and organisations have been attracted by the use of questionnaires completed by patients and colleagues as a means of obtaining multisource feedback on the performance of individual doctors. While aiming to reflect what a doctor actually does in clinical practice,[5] questionnaire responses may have shortcomings of their own. In a critique of instruments for the rating of doctors by their colleagues, Evans *et al*[6] identified a dearth of published information concerning the theoretical framework underpinning them and a lack of evidence of construct and criterion validity supporting their use. A recent review[7] of 10 questionnaires designed to gather feedback from patients concluded that few have undergone rigorous testing of reliability and validity.

Despite these shortcomings, the first hand experience which colleagues and patients have of the way in which doctors perform make the use of suitably validated questionnaires attractive as a component of the revalidation process. Recent years have seen an increase in the published evidence informing the appropriate use of multisource feedback in the evaluation of doctors' professional performance.[8–16] Survey instruments used for this purpose need to have a sound theoretical basis and to cover all of the relevant domains.

In 2004 the GMC's registration committee commissioned a review of existing questionnaires and found that none of them covered all of the necessary domains adequately. A working party therefore devised patient and colleague questionnaires specifically for use in revalidation, building on earlier work in the field.[8 17] Unlike most existing questionnaires, the GMC questionnaires were primarily summative in intent, although it was anticipated that the results would be fed back to

doctors as a basis for reflection and, if appropriate, remediation. The face validity of these questionnaires was established by Market and Opinion Research International (MORI) through a series of focus groups,[18] and preliminary assessment of the properties of the questionnaires was undertaken by the University of Leeds.[19]

It is of vital importance that patients, employers, governments and doctors have confidence in the processes and instruments adopted in the revalidation of doctors. The UK Postgraduate Medical and Education Training Board (PMETB) has stated that the reliability of survey instruments for use in workplace assessment of doctors is of "central importance" and that "if [such] a test is not reliable it cannot be valid".[20] Here we report on the findings of a survey investigating the utility of the GMC patient and colleague questionnaires in assessing the professional performance of a large sample of UK doctors.

METHODS

Survey instruments

The GMC survey instruments comprise two questionnaires, each presenting questions relating to the seven domains of *Good medical practice*,[21] the GMC's core guidance on the principles and values to which it requires registered doctors to adhere. Each questionnaire (table 1) is prefaced by a brief introduction and explanation of the purpose of the survey. The patient questionnaire comprises 18 items (3 contextual; 11 performance evaluation; 3 descriptive of the respondent; and 1 free-text) and the colleague questionnaire comprises 25 items (2 contextual; 18 performance evaluation; 4 descriptive of the respondent, 1 free-text item). Nine patient questionnaire items and 17 colleague questionnaire items invite responses on a five-point Likert scale with descriptives of: poor (scoring 1), less than satisfactory, satisfactory, good and very good (scoring 5); or strongly disagree (scoring 1), disagree, neutral, agree and strongly agree (scoring 5). Two patient questionnaire items and one colleague questionnaire item invite binary (yes/no) performance evaluation responses.

Sampling and recruitment of doctors

Prior to undertaking the main survey, a pilot study was conducted with 46 volunteer doctors to inform recruitment and survey process issues. Data from these doctors were not combined with the main survey data set. For the main survey, we approached a convenience sample of doctors from a range of National Health Service (NHS) settings across the UK (England, Northern Ireland, Scotland, Wales), practising in communities with socioeconomically and culturally diverse profiles. We sampled doctors from acute and primary care settings, inviting the contribution of established and training grade doctors, but excluding doctors from specialties not routinely involved in face-to-face patient consultations. Following research governance approval, we sampled doctors from 5 acute and 11 primary care trusts, 1 mental health trust, and general practice (GP) registrars from a deanery. In addition we invited doctors from a range of non-NHS settings to participate, including prison doctors, occupational physicians, out-of-hours primary care medical practitioners, locum doctors, independent practitioners, and doctors undergoing GMC performance review. Our target sample size of doctors was not based on a formal sample size calculation, but rather aimed to obtain a large sample with sufficient data to permit an assessment of the performance of the questionnaires in line with recognised best practice.[22] We aimed for substantially more than 150 respondents across a range of practice settings and clinical specialties to allow for reliable estimation of correlation coefficients and analysis of principal components.[23 24]

The approach to sampling was pragmatic—we sampled a substantial proportion of doctors in each participating trust setting. Not all trusts provided us with age and gender information of doctors in the sampling frame, and no sampling frame was readily identifiable for doctors working in the non-NHS settings.

Administering the colleague questionnaire

Doctors who initially agreed to participate were invited to complete and return to the research team a list of up to 20 colleagues who would be in a position to comment on their professional practice and/or behaviour. We advised participants that approximately half the colleagues should be medical peers, with the other half drawn from other occupations related to healthcare. Based on findings from the pilot study and international evidence,[9 17] we aimed to secure at least 12 completed colleague questionnaires for each doctor. Specific instructions for identifying colleagues were provided to participants.

We then approached each doctor's nominated colleagues by email, inviting them to complete an online questionnaire addressing the professional behaviour and practice of the named doctor. Colleagues were requested to complete the questionnaire within 2 weeks, and were provided with an information sheet, and a security PIN number to access the questionnaire.

Table 1 Structure of the patient (PQ) and colleague (CQ) questionnaires

Item domains	PQ number of items (item number in PQ)		CQ number of items (item number in CQ)	
Introduction/explanation	+		+	
Contextual items				
Reason for attendance	2	(1, 2)	–	–
Previous experience of this doctor	1	(7)	–	–
Professional role and frequency of contact with doctor	–	–	2	(22, 23)
Performance evaluation items*				
Generic assessment	9	(3a–g, 4a,b)	17	(1–17)
Global assessment	2	(5, 6)	1	(18)
Participant descriptive items				
Demographics/ethnicity	3	(8–10)	4	(19–21, 24)
Free text	1	–	1	–
Total	18		25	

*For details of individual items, see copies of questionnaires online.

One email reminder was sent to non-responders after 2 weeks. Paper completion was available to colleagues on request.

Administering the patient questionnaire

The patient questionnaire was distributed as a post-consultation "exit" survey. Doctors were provided with 45 consecutively numbered patient questionnaires to be distributed by administrative staff to a consecutive sample of patients for completion after seeing the doctor. Patients were asked to place the completed questionnaire in a sealed envelope and return it to a collection point. Informed by evidence from similar studies[8 11 25] and our pilot study, we aimed to obtain 30 completed patient questionnaires for each doctor. Doctors working in out-of-hours settings, where a substantial proportion of their consultations may be over the telephone, used a different standardised procedure to capture responses from patients following telephone consultations, treatment centre attendance or home visits. Doctors initially volunteering to participate but not returning any patient questionnaires were sent reminders at 4 and 12 weeks.

Informal feedback on the survey process

During the recruitment and fieldwork, we encouraged doctors to provide informal feedback on the survey processes. Field notes were collated by study administration staff, or through written feedback from participants.

Data management

Questionnaires were checked for discrepant or invalid responses and any missing data were identified. Patient questionnaires were designed to be electronically scanned, although manual data entry was used for any questionnaire where scanning proved impossible.

Data analysis

Descriptive statistics were used to characterise doctors who participated in the study and the patients and colleague respondents who completed questionnaires. Where possible, participant doctors were compared with non-participants with respect to demographic data. The analysis of the free text comments is not reported here. The acceptability of the questionnaires was investigated through an examination of overall and individual item response rates. The reliability of the questionnaires was evaluated at respondent level, and also at participant doctor level where, in line with normal practice,[8 9 11 12 17 26] mean scores for professional performance items were examined. The proportion of responses in the lowest two of five response categories was compared for each item in the two questionnaires. Principal components analysis was used to investigate the relationship between responses to each item in each of the two questionnaires. We explored the potential for rotation of resulting factors in improving their interpretation. The Spearman Brown prophecy formula[22] was used to estimate the minimum number of completed patient and colleague questionnaires required per doctor to achieve reliable estimates of doctor performance.

Two measures of each doctor's performance were derived based on the sum across the nine generic assessment items in the patient questionnaire (patient questionnaire score) or the 17 generic assessment items in the colleague questionnaire (colleague questionnaire score). Standardised versions[27] of these scores were calculated, based on the sum across all items of the loading for the item multiplied by the doctor's mean score for the item. Differences between male and female doctors in respect of patient-derived and colleague-derived scores were investigated by comparing mean scores, and the association of scores with the doctor's age was investigated using the Spearman rank correlation coefficient. Where a doctor had returned both patient and colleague questionnaire data, inspection of a scatterplot and the Pearson correlation coefficient was used to investigate the relationship between the two standardised scores. Post hoc comparisons (Tukey test) were used to investigate the impact of clinical setting on both patient-derived and colleague-derived scores for individual doctors.

We adopted a norm-referenced approach to standard setting, identifying doctors who had a z score of less than −1.96 (ie, doctors whose scores fell approximately 2 SD below the cohort mean) on either the patient or colleague summary scores as outliers.

Generalisability theory

Although this is a naturalistic dataset drawn from "real life" clinical settings, a decision (D) study was conducted to investigate the contribution to the overall variance by different numbers of patient or colleague raters providing judgements against the nine items in the patient questionnaire or the 17 items in the colleague questionnaire, and with the object of measurement as the index doctor being assessed. We used a random effects model in each of the datasets.

RESULTS

Describing study requirement and participation

Eighteen NHS trusts identified a total of 2589 doctors, who formed a sampling frame for the study; 450 of these doctors (17.4%) gave preliminary agreement to participate in the study. The proportion of doctors in trusts invited to participate ranged between 34.0% and 100% (mean (SD) 87.1% (30.0%)). Another 91 doctors from a variety of other NHS and non-NHS settings agreed to provide data. By the close of data collection, a maximum of 9 months after the initial approach, 398 doctors (73.6% of those giving preliminary agreement) had returned some data. Both colleague and patient questionnaires had been returned by 291 doctors. Although this was not formally assessed, many doctors and other individuals involved in the survey commented favourably on the ease of the processes involved.

Eight of the 18 participating trusts provided us with complete or near-complete (≥95%) information on the age and gender of doctors within their organisation. Of 1133 doctors approached in such settings, 212 (18.7%) agreed to take part. No response was obtained from 764 (67.4%) doctors, and 157 (13.9%) declined to contribute. In these settings, the age and gender profile of doctors who participated in the study was similar to that of doctors who did not participate in the study (mean age in years = 45.7 (8.0) vs 44.6 (9.6), $t = 1.42$, $p = 0.157$; 71/211 (33.6%) vs 279/913 (30.6%) female, $\chi^2 = 0.764$, 1 df, $p = 0.113$).

From initial pilot work undertaken with 46 doctors[28] we estimated that:

- the median (interquartile range (IQR)) time between sending the patient questionnaire pack to participant doctors and receiving complete data was 56.5 (43.0, 70.0) days;
- approximately 80% of patients who were offered a post-consultation "exit" survey would accept and complete it;

- the colleague questionnaire data collection process took a median of 74.0 (58.75, 89.25) days from initial contact with the doctors to the completion of the CQ data set.

Of this time, 41 days delay was incurred in waiting for completion and return of the colleague list. For logistical reasons we could not replicate the analysis of the time taken to return patient or colleague responses during the course of the main survey. However, our experience suggests that the initial estimates observed in the pilot were broadly representative of our experiences during the main survey.

Patient survey

Responses were obtained from 13 754 patients attending one of 380 doctors (median response per doctor 37 (IQR 31, 42)). The median age group of patient respondents was 41–60 years, with a preponderance of female respondents among those identifying their gender (7069/11 939, 59.2%). Most respondents identifying their ethnic status were "white British" (n = 11 777/13 100, 89.9%). The questionnaire appeared acceptable to patients, with only 365 (2.7%) and 85 (0.6%) completing <50% or <10% of all items, respectively. Missing data on performance ranged from 211 (1.6%) to 772 (5.6%) per item. Six items had <2% of respondents reporting that the item content "did not apply", whereas the remaining three items (explanation of condition and treatment, involvement in decision making, and the provision or arrangement of treatment) had 475 (3.5%), 922 (6.7%), and 1316 (9.6%) of respondents, respectively, reporting that the option did not apply. Respondents omitting data on age, gender or ethnicity represented 1815 (13.2%), 1698 (12.3%) and 651 (4.7%) of the sample respectively.

Patient responses were highly skewed towards favourable impressions of doctor performance, with mean (SD) scores out of a maximum possible score of five across each of nine performance evaluation items ranging from 4.68 (0.66) to 4.88 (0.42). On the two items requiring yes/no answers, 13 341/13 415 (99.4%) of respondents were confident of the doctor's ability to provide care, and 10 214/12 982 (78.7%) had no reservations about seeing the same doctor again. Across the nine performance evaluation items, the reliability coefficient (Cronbach α) was 0.898 with an average inter-item correlation coefficient of 0.526 (range 0.260–0.855). The average score (range) across the nine items was 4.80 (4.70–4.88) out of a maximum possible score of 5.00. Further psychometric data relating to the patient questionnaire is provided online (web only table A).

The question "I have no reservation about seeing this doctor again", answered using binary response categories, appeared to cause confusion among patient respondents, with 87 (0.6%) respondents corrupting the item by altering their response category and/or word stem to clarify their response. Furthermore, this item had a substantially higher proportion of adverse ratings (2768/12 982, 21.3%) compared with the two items with the next highest proportion (1.6%) of adverse ratings (relating to doctor confidentiality and honesty/trustworthiness).

Principal components analysis of individual patient responses to the nine performance evaluation items with orthogonal (varimax) rotation of resulting factors identified two components (web only table B), together accounting for 76.8% of the variance in the sample. The first component comprised the first seven of the performance evaluation items (patient questionnaire items 3a–g (see web only fig A); loadings 0.810–0.851), whereas the second component comprised the last two of the performance evaluation items (patient questionnaire items 4a, b (see web only fig A); loadings 0.946–0.951).

Colleague survey

The median (IQR; range) number of colleagues nominated by participating doctors was 20 (18, 20; 3–26), for each participating doctor. Responses were obtained from 4269 colleagues relating to 309 doctors (median (IQR) 14 (12, 17) responses per doctor). Most colleague questionnaires were completed using the online questionnaire (3363/4269, 78.8%), whereas the remainder (906, 21.2%) were manually entered following receipt of a paper questionnaire. The mean (SD) age of colleague respondents was 45.3 (8.78) years and 2485/4248 (58.5%) were women. The majority of colleagues (3603/4171; 86.4%) reported their ethnic group as white British. Consistent with our instructions to participants, approximately half of colleague respondents were doctors (2107/4236, 49.7%); those describing their role as registered nurse (n = 754) comprised 17.8% of the sample and the remainder (n = 1377, 32.5%) adopted a wide range of role descriptors including allied healthcare professional, healthcare assistant, practice manager, administrator, pharmacist or one of around 300 free text descriptors submitted by participants. Sociodemographic data were not available for non-respondents. Only 71/4269 (1.7%) of colleague respondents completed <50% of the questionnaire items.

Colleague responses were highly skewed towards favourable impressions of doctor performance. Mean (SD) scores out of a maximum possible score of 5 for each of 17 performance evaluation items ranged from 4.49 (0.68) to 4.87 (0.38) with clear evidence of a ceiling effect (41.7–87.1% of respondents using the highest available category across the 17 items; only 0.1–0.7% used either of the lower two response categories). Missing data varied between 0.1% (confidence in doctor's practice of confidentiality) and 2.9% (supervision of colleagues). However, larger numbers of colleagues (range 1.5–26.0%) indicated that they did not have knowledge of certain aspects of the doctor's performance, most notably in relation to teaching (26.0%) and supervising colleagues (25.5%).

The reliability of the colleague questionnaire was good (Cronbach α for the 17 performance evaluation colleague questionnaire items was 0.922) with an average inter-item correlation of 0.418 (range 0.189–0.725). The mean score (range) across the 17 items was 4.71 (4.50–4.90) out of a maximum possible score of 5.00. Further psychometric data relating to the colleague questionnaire is provided online (web only table C).

Principal components analysis identified three components in the 17 colleague questionnaire items, which together accounted for 61.0% of the total variance in the sample. Inspection of the scree plot (fig 1) identified the substantial dominance of the first component with loadings (range 0.535–0.780, web only table D) of all 17 items (items 1–17, web only fig B) on the first component.

Determining the minimum sample size required

Application of the Spearman Brown prophecy formula identified that acceptable reliability[29] (α >0.85) was achieved with a minimum of 22 completed patient questionnaires or 8 completed colleague questionnaires per doctor. Of 309 doctors who had colleague data returned, 288 had ⩾8 colleague responses, and of 380 doctors who returned patient data, 355 returned ⩾22 patient questionnaires. Overall, 252/291 (86.6%) doctors returning both patient and colleague questionnaires returned sufficient numbers of questionnaires for both surveys.

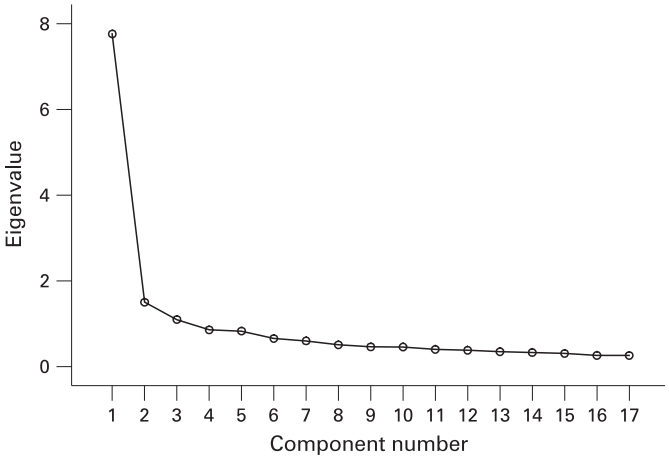

Figure 1 Scree plot for 17 performance evaluation colleague questionnaire items (colleague level analysis).

Twenty-three patient ratings resulted in a G coefficient (standard error of measurement) of 0.65 (0.16), whereas 36 patient ratings resulted in G = 0.75 (0.13). Seven (7) colleague ratings resulted in G = 0.65 (0.25) and 12 colleague ratings resulted in G = 0.76 (0.19).

Analysis of doctor-level scores

The mean (SD) item values across 17 colleague questionnaire performance evaluation items ranged from 4.45 (0.37) to 4.86 (0.17). For nine patient questionnaire performance evaluation items, the mean (SD) item values ranged from 4.68 (0.17) to 4.87 (0.14). The mean (SD) summed scale scores for colleague and patient questionnaires were 79.35 (3.46) out of a maximum achievable of 85.00 and 43.01 (1.33) out of a maximum achievable of 45, respectively. Cronbach α for the 17 colleague mean item scores was 0.947, with a mean (range) inter-item correlation among the 17 items of 0.52 (0.23–0.88). For the nine performance evaluation items in the patient questionnaire, the reliability (α) was 0.962 with an mean (range) inter-item correlation of 0.53 (0.26–0.86).

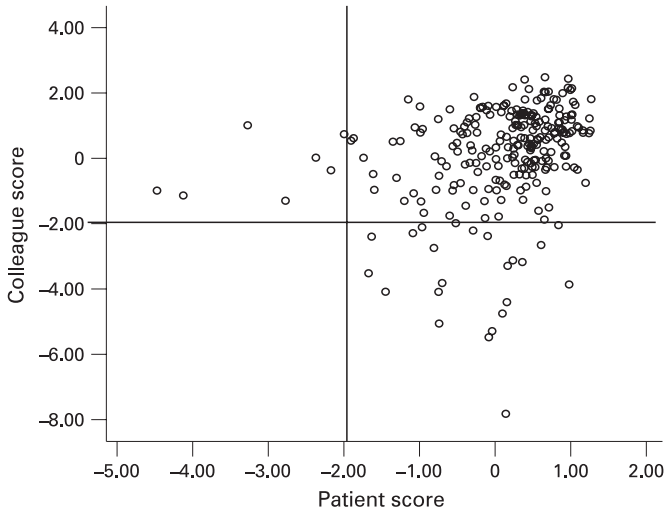

Figure 2 Patient and colleague scores (standardised measures) for 252 doctors with ⩾22 patient questionnaires and ⩾8 colleague questionnaires. Internal reference lines added at z < −1.96.

Outlying performance

Patient and colleague scores for each doctor were only moderately[30] correlated (Spearman ρ = 0.344, p<0.01). In this volunteer sample of doctors, 7/252 (2.8%) doctors had patient standardised mean scores with z < −1.96, and 22 (8.7%) had colleague standardised mean scores with z < −1.96. No doctor had z values < −1.96 for both patient and colleague scores (fig 2). The processes of identifying outlying doctors was sensitive to the sample size available (web only table E) with doctors achieving small samples more likely to be identified as outliers. There was no significant difference between male and female doctors in respect of patient and colleague questionnaire scores, but older doctors had lower scores in respect of both questionnaires compared with younger doctors (Spearman ρ = −0.107, p = 0.05 and ρ = −0.219, p<0.001 for patient and colleague scores, respectively).

As closer inspection of the distribution of the differences in standardised mean scores between settings for patient and colleague ratings revealed significant homogeneity of the variances (Levene statistic = 2.856 and 3.301, respectively), we adopted a significance level 0.025 when interpreting the post hoc comparisons.[23] Doctors from different settings differed in respect of patient scores, but not on colleague scores. Patient-derived scores were similar for doctors working in primary care and acute trusts, and "other" settings. Doctors from the mental health trust had lower patient scores compared with doctors from acute trusts (mean difference in z score (SE) = −0.910 (0.280), p = 0.007) and from primary care trusts (−0.977 (0.275), p = 0.002), but comparable scores with those from other settings (−0.527 (0.327), p = 0.372).

DISCUSSION

This study has demonstrated the utility of the GMC patient and colleague questionnaires in collecting evidence regarding the professional performance of doctors. The UK government has recently stated that multisource feedback of this type will in future comprise an element in the annual appraisal of doctors working in the NHS and it clearly has potential for use in the process of revalidation.[31]

Doctors from a wide range of clinical practice settings in the UK contributed to the study, supporting the generalisability of the findings. Where data were available, the sociodemographic characteristics of participant doctors were broadly comparable with non-participants, suggesting that recruited doctors were broadly representative of doctors working in NHS trusts/organisations as a whole. Notwithstanding this, the doctors contributing to this study were volunteers and may not be representative of all doctors in terms of their interactions with patients and colleagues. Indeed, there is evidence to suggest that, within primary care, practices which involve themselves in research differ from other practices in respect of certain features (such as the use of generic prescribing) which have also been associated with differences in quality of care.[32] Although achieving a diverse sample, we did encounter some difficulty in securing participation from doctors working in non-NHS settings. Informal feedback (not reported here) suggested that such doctors may have been uncertain about the relevance of the survey process and the content of questionnaire items to their particular clinical environments. Similarly, following completion of the study, some doctors raised concerns about the nature of feedback provided by patients who were being seen in what were perceived as challenging clinical environments—for example, doctors providing care in prisons,

emergency departments and mental health settings or working in occupational health.

Analysis of the psychometric properties of the patient and colleague surveys showed that both surveys were acceptable to patients and colleagues. Most doctors were able to achieve sufficient numbers of patient and colleague responses, and the data within the questionnaires demonstrated low levels of missing data for individual performance evaluation items. Although around 1 in 4 colleague respondents was "unable to comment" on key aspects of a doctor's professional practice (chiefly teaching and supervision of colleagues), we propose that these items are retained on account of their contribution to the content specificity and comprehensiveness of the colleague survey instrument in addressing key principles of professional behaviour laid out by the GMC.[21] It is of note that a similar Canadian instrument[8] had low levels ("less than 10%") of "unable to comment" responses, but did not include items relating to either of the domains to which we have referred. One item from the patient survey ("I have no reservation about seeing this doctor again"; binary response yes/no) did, however, cause respondents some confusion, and we recommend that the wording of this item is revised.

Using a classical approach to establishing reliability,[22] both patient and colleague questionnaires were highly reliable measures, demonstrating high internal consistency with α scores comparable with or exceeding other questionnaires having similar intent.[8 17] Some researchers[33 34] have called for the development of new and more sophisticated approaches in the assessment of clinical performance. With the same number of patient or colleague raters, the generalisability D-study produced lower coefficients than Cronbach α measure of internal consistency. Given that generalisability theory takes into account the contribution of various sources of error and not only the internal inconsistency of the questionnaires items affecting the true score (ie, the doctor's professional behaviour), this difference is to be expected. These findings emphasise the naturalistic study design and setting. The use of untrained patient and colleague raters seems inevitable in the use of multisource feedback in routine clinical settings where the intention is to obtain information from sources reflecting routine clinical care. Although training of assessors might result in improved generalisability ratings, the process of training may in itself undermine the attempt to capture information on what the doctor does[5] in routine clinical practice.

In the UK, the PMETB has proposed[20] that high stakes assessments should have a reliability (Cronbach α) of at least 0.9. Both of these questionnaires meet that criterion, and have α coefficients broadly in line with internationally recognised and adopted instruments of similar intent.[8 11] The G coefficients associated with these questionnaires are also in line with other internationally recognised instruments having similar intent—for example the American Board of Medical Specialties (ABIM) patient questionnaire has a mean (SD) score of 4.8 (0.13) on a five-point scale, and a G coefficient (95% CI) of 0.67 ($\pm$0.14).[11] The ABIM peer questionnaire has a mean (SD) score of 7.9 (0.34) on a nine-point scale with a G coefficient of 0.61 ($\pm$0.41).[11] Similar equivalence is presented in evidence emanating from Canada[26] (patient instrument: G = 0.71 for 25 patients using a 13-item survey, G = 0.61 for 8 colleagues using a 21-item questionnaire).

The validity of the patient questionnaire was supported by the process of questionnaire development involving preliminary qualitative work undertaken with patients and an initial investigation of the properties of the questionnaires. Furthermore, we observed the consistency of our findings with other studies[35] in respect of the more favourable impressions of doctor performance expressed by older as compared with younger patients, and by "white British" respondents as compared with those from ethnic minorities.

While we aimed to analyse at least 30 patient questionnaires and at least 12 colleague questionnaires for each doctor, our data suggest that for these instruments acceptable internal consistency is achieved by the return of completed questionnaires from 22 patients and eight colleagues. Whilst this is consistent with other studies of patient[11 26] and peer[9 12 17 36 37] feedback we feel that our conclusion must be regarded as tentative until information from a larger and more comprehensive sample of UK doctors is available. Given the sensitivity of patient-derived scores to the clinical setting, there is a need to obtain specialty-specific benchmark data using surveys of patients and colleagues of a larger numbers of doctors within particular specialties and in primary care settings. Further research is also required to inform patient and colleague sampling strategies, for example, to explore variation between groups of health professionals sharing similar training or qualifications, or patients derived from differing social or demographic groups.

Using a predetermined norm-referenced definition of outlying performance, we observed no overlap between the doctors identified as outliers by patients and those identified as outliers by colleagues. Other studies have reported similar findings,[38 39] and it thus appears that patient-derived and colleague-derived scores reflect differing and complementary aspects of doctor's performance. In line with the experience of others,[40 41] older doctors were observed to have lower performance scores when compared with the scores of younger doctors.

The observation that doctors identified as outliers returned fewer questionnaires than doctors who were not outliers has practical importance to the establishment of robust survey processes. It will be necessary to ensure that doctors using the questionnaires comply with minimum survey sampling requirements. We cannot comment on the possibility that colleagues with adverse views of a doctor's performance declined to participate in the study when invited to do so, but note recent authoritative guidance[21] highlighting the importance of colleague feedback, and advising UK doctors that they should be willing to contribute honest and objective assessments of colleagues.

Limitations

Of necessity, participants in the study were volunteers. It seems likely that those who agreed to participate were reasonably confident about their own standards of practice and the sample may have been skewed towards good performance. This makes it difficult to draw conclusions about the sensitivity of the questionnaires in detecting doctors whose performance merits further scrutiny. The questionnaires investigated in this study have mean scores and distributions very similar to a range of other internationally accepted and recognised patient[8 11 42] and colleague[8 9 11 17 42] instruments, which are clearly skewed towards favourable impressions of doctor performance. Whether, as suggested by some authorities,[37 43] the skewed responses may be in part due to the perceived purpose of the survey cannot be determined but should be considered in interpreting the results. Data derived from use of the questionnaires in larger, unselected groups of doctors will need to be scrutinised to allow refinement of the criteria used to identify outlying performance.

Although participant doctors were asked to distribute the survey to consecutive patients being seen, we did not check that this had taken place. We invited participant doctors to identify colleagues who might be in a position to comment on their professional behaviour and practice on the basis of evidence suggesting that such assessments are not substantially affected by the method of peer-assessor selection[17]; but further work in respect of the most appropriate processes for colleague assessor identification is warranted, along with an assessment of the temporal stability of responses (test–retest reliability) of the questionnaires.

CONCLUSIONS

Patient and colleague surveys have potential as a means of collecting information regarding doctors' performance. Although both patient and colleague data were skewed towards favourable impressions of performance, the approach outlined here enabled us to discriminate between doctors in respect of their professional performance. The lack of overlap between doctors identified as outliers by patient and colleague scores suggests that patients and colleagues provide independent perspectives on doctor performance, and that both sources of feedback are required.

Given the volunteer nature of the sample and the use of norm-referenced approaches to standard setting, we would urge caution about identifying any of the doctors who participated in the study as displaying deficient performance. The GMC considers that questionnaires might be used as one of several methods of identifying doctors whose practice requires further scrutiny but not as an absolute and free standing measure of performance. Further validation surveys within the context of a census sample of doctors undergoing revalidation are needed to establish precise criteria that would trigger such further scrutiny and also to determine whether satisfactory questionnaire scores are reliable indicators of acceptable professional performance.

Acknowledgements: We are grateful to all the patients, doctors and their colleagues who contributed to this study. Professor J McLachlan (University of Durham, UK) and Professor LSchuwirth (University of Maastricht, the Netherlands) contributed to the development of the questionnaires, and advised the GMC on the design of the study. Dr C Ricketts and Dr J Archer (Peninsula Medical School, UK) and Dr G Ponnaperuma (University of Dundee) provided valuable feedback and statistical advice regarding data interpretation and analysis, and provided useful comments on the text of the paper.

Funding: The UK General Medical Council funded this study.

Competing interests: MG is a director of Client Focussed Evaluation Programme (UK). SB is former chair of the UK GMC registration committee (to Jan 2008).

Ethics approval: Agreement to this study was provided by the chair of the North and East Devon NHS Research Ethics Committee.

REFERENCES

1. **Department of Health, Chief Medical Officer**. *Good doctors, safer patients—proposal to strengthen the system to assure and improve the performance of doctors and to protect the safety of patients*. London: Department of Health, 2006:1–202.
2. **General Medical Council, Irvine D**. *Maintaining good medical practice*. London: General Medical Council, 1998.
3. **Norcini JJ.** Psychometric issues in the use of practice performance assessment for physician evaluation. In: Mancall EL, Bashook PG, eds. *Evaluating residents for health board certification*. American Board of Medical Specialties. Washington DC: Evanston IL, 1998.
4. **Norcini JJ.** Recertification in the United States. *BMJ* 1999;**319**:1183–5.
5. **Miller GE.** The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65**:S63–7.
6. **Evans R,** Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004;**328**:1240–3.
7. **Chisholm A,** Askham Janet, Picker Institute Europe. *What do you think of your doctor? A review of questionnaires for gathering patients' feedback on their doctor*. Oxford: Picker Institute Europe, 2006:1–58.
8. **Hall W,** Violato C, Lewkonia R, *et al*. Assessment of physician performance in Alberta: the physician achievement review. *CMAJ* 1999;**161**:52–7.
9. **Archer JC,** Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005;**330**:1251–3.
10. **Archer J,** Norcini J, Southgate L, *et al*. Mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? *Adv Health Sci Educ Theory Pract* 2008;**13**:181–92 [Epub 12 Oct 2006].
11. **Lipner RS,** Blank LL, Leas BF, *et al*. The value of patient and peer ratings in recertification. *Acad Med* 2002;**77**:S64–6.
12. **Whitehouse A,** Hassell A, Wood L, *et al*. Development and reliability testing of TAB a form for 360 degrees assessment of senior house officers' professional behaviour, as specified by the General Medical Council. *Med Teach* 2005;**27**:252–8.
13. **Lockyer JM,** Violato C, Fidler H. A multi source feedback program for anesthesiologists. *Can J Anaesth* 2006;**53**:33–9.
14. **Violato C,** Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ* 2003;**326**:546–8.
15. **Violato C,** Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. *Pediatrics* 2006;**117**:796–802.
16. **Norcini JJ.** Current perspectives in assessment: the assessment of performance at work. *Med Educ* 2005;**39**:880–9.
17. **Ramsey PG,** Wenrich MD, Carline JD, *et al*. Use of peer ratings to evaluate physician performance. *JAMA* 1993;**269**:1655–60.
18. **MORI Social Research Institute**. *Revalidation questionnaires testing: qualitative research findings*. London: General Medical Council, 2004.
19. **Kilminster S,** Pell G, Roberts T. Patient and colleague questionnaires: validation report to the GMC. Leeds: University of Leeds, Medical Education Unit, 2005.
20. **Postgraduate Medical Education and Training Board**. *Developing and maintaining an assessment system—PMETB guide to good practice*. London: Postgraduate Medical Education and Training Board, 2007.
21. **General Medical Council (Great Britain)**. *Good medical practice*. London: General Medical Council, 2006.
22. **Streiner DL,** Norman GR. *Health Measurement Scales: A practical guide to their development and use*. Oxford: Oxford University Press, 2003.
23. **Tabachnick BG,** Fidell LS. *Using multivariate statistics*. New York: HarperCollins College Publishers, 1996.
24. **Guadagnoli E,** Velicer WF. Relation of sample size to the stability of component patterns. *Psychol Bull* 1988;**103**:265–75.
25. **Wensing M,** van deVleuten C, Grol R, *et al*. The reliability of patients' judgements of care in general practice: How many questions and patients are needed? *Qual Health Care* 1997;**6**:80–5.
26. **Lockyer J,** Blackmore D, Fidler H, *et al*. A study of a multi-source feedback system for international medical graduates holding defined licences. *Med Educ* 2006;**40**:340–7.
27. **Altman DG.** *Practical statistics for medical research*. London: Chapman & Hall, 1991.
28. **Campbell JL,** Richards S, Dickens A, *et al*. The GMC patient and colleague questionnaire study: report of the pilot study. Exeter: Peninsula Medical School, 2006.
29. **Weiner EA,** Stewart BJ. *Assessing individuals: psychological and educational tests and measurements*. Boston: Little, Brown, 1984.
30. **Cohen J.** *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
31. **Department of Health**. *Trust, assurance and safety—the regulations of health professionals in the 21st century*. London: Stationery Office, 2007:5–12 (White Paper.)
32. **Hammersley V,** Hippisley-Cox J, Wilson A, *et al*. A comparison of research general practices and their patients with other practices—a cross-sectional survey in Trent. *Br J Gen Pract* 2002;**52**:463–8.
33. **Schuwirth LW,** van der Vleuten CP. A plea for new psychometric models in educational assessment. *Med Educ* 2006;**40**:296–300.
34. **Crossley J,** Davies H, Humphris G, *et al*. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;**36**:972–8.
35. **Campbell JL,** Ramsay J, Green J. Age, gender, socio-economic, and ethnic differences in patients' assessments of primary health care. *Qual Health Care* 2001;**10**:90–5.
36. **Violato C,** Marini A, Toews J, *et al*. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med* 1997;**72**:S82–4.
37. **Norcini JJ.** Peer assessment of competence. *Med Educ* 2003;**37**:539–43.
38. **DiMatteo MR,** DiNicola DD. Sources of assessment of physician performance: a study of comparative reliability and patterns of intercorrelation. *Med Care* 1981;**19**:829–42.
39. **Joshi R,** Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. *Acad Med* 2004;**79**:458–63.
40. **National Patient Safety Agency**. *National Clinical Assessment Service: analysis of the first four years' referral data*. London: National Patient Safety Agency, 2006.
41. **Choudhry NK,** Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. *Ann Intern Med* 2005;**142**:260–73.
42. **Wood J,** Collins J, Burnside ES, *et al*. Patient, faculty, and self-assessment of radiology resident performance: a 360-degree method of measuring professionalism and interpersonal/communication skills. *Acad Radiol* 2004;**11**:931–9.
43. **Hay JA.** Tutorial reports and ratings. In: Shannon S, Nocterm G, eds. *Evaluation methods: a resource handbook*. Hamilton, Ontario: McMaster University, 1995.