GMC Multi-Source Feedback Study

Scientific report of the Main Survey (2008-10)

Executive Summary

Professor John Campbell (Academic Lead) [1.]

Ms Jacqueline Hill (Research Fellow) [1.]

Dr Jeremy Hobart (Reader) [2.]

Professor Ajit Narayanan (Statistician) [3]

Professor Geoff Norman (Consultant) [4.]

Dr Suzanne Richards (Senior Lecturer) [1.]

Mr Martin Roberts (Statistician, Research Fellow) [1]

Dr Christine Wright (Research Fellow) [1.]

1. Primary Care Research Group, Peninsula College of Medicine & Dentistry, Smeall Building, St Luke's Campus, Exeter, EX1 2LU.
2. Peninsula College of Medicine & Dentistry, Room N13, Tamar Science Park, Davy Road, Plymouth, PL6 8BX.
3. School of Computing & Mathematical Sciences, Auckland University of Technology, 2-14 Wakefield Street, Auckland 1142, New Zealand.
4. Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton Ontario, Canada.

Executive Summary prepared: 10 February 2012

PENINSULA
COLLEGE OF MEDICINE & DENTISTRY
UNIVERSITIES OF EXETER & PLYMOUTH

Overview

This Executive Summary reports the findings of a large-scale survey commissioned as part of a three year programme of research funded by the UK General Medical Council (GMC). The primary aim of the programme was to evaluate a survey of doctors undertaking multi-source feedback (MSF) using the GMC patient and colleague questionnaires. A series of confidential reports were prepared between November 2010 and March 2011 for review by the GMC and their technical advisers. It is the academic team's intention to seek peer review and publication of the data in the scientific literature.

In this Executive Summary, we provide an overview of doctor recruitment and data collection (as at 17 January 2011), together with a summary of the scientific analysis of the data collected from both patients and colleagues. We report on a range of statistical approaches that we have used to extend our understanding of the performance of the GMC questionnaires.

A series of sub-studies and sub-analyses were conducted in parallel to the main survey work to examine specific questions in relation to the implementation of MSF in applied settings. The findings of these studies have not been included in this document, but we aim to seek publication of these in relevant peer-reviewed academic journals in due course.

Whilst this Executive Summary is the output of the Peninsula College of Medicine and Dentistry research team, we acknowledge the extensive support of our partner organisation, CFEP-UK Surveys Ltd, who were commissioned by the GMC to administer the data collection. In particular, Associate Professor Michael Greco has been instrumental in the development of the work from its earliest stages.

Professor John L Campbell
Peninsula College of Medicine & Dentistry, Exeter
10 February 2012

Contents

1 Background

All doctors who wish to practise medicine in the United Kingdom need to be registered with and licensed by the General Medical Council (GMC). To retain their licence with the GMC, doctors will also be required to revalidate on a regular basis to demonstrate that they are still 'fit to practise' medicine.

The GMC's proposals for the revalidation process have now been published.[1-3] As part of that process, doctors will need to collate 'supporting information'[4] to demonstrate they meet relevant standards of professionalism. Such information will feed into a 'strengthened' appraisal process,[3;4] and ultimately lead to a recommendation by a Responsible Officer to the GMC about the doctor's suitability for revalidation. The collection of feedback from colleagues and patients, using structured questionnaires is likely to be an integral component of the revalidation process for doctors.[4;5]

Early reviews[6;7] of the tools available for this purpose (conducted in 2004 and 2006) suggested that few had undergone sufficiently rigorous testing of reliability and validity. In the absence of suitably robust instruments to assess the professional performance of doctors, the GMC developed its own patient and colleague questionnaires, with content based on the principles of 'Good Medical Practice'.[8] A series of pilot studies have explored their utility, reliability and validity.[9-11] These studies suggested the questionnaires were acceptable to respondents, offered preliminary reassurance regarding their validity, reliability and feasibility, and indicated they could identify a range of performance in respect of professionalism in a volunteer sample of doctors. However, the test-retest reliability and convergent validity of the GMC questionnaires had not been established.

In 2010, a report commissioned by the Royal College of General Practitioners[12] reviewed nine colleague feedback instruments and ten patient feedback instruments. That report concluded that considerable work had been undertaken to develop

questionnaires that might be capable of assessing the professional performance of doctors for the purposes of revalidation. In particular, the review noted that the GMC Patient Questionnaire (PQ) and the GMC Colleague Questionnaire (CQ) mapped well onto the competencies required for 'Good Medical Practice'[8] and that comprehensive psychometric testing, in line with the GMC's Framework Attributes,[3] had been undertaken.

The GMC questionnaires have recently undergone further development and evaluation using a large sample of UK doctors working across a wide range of clinical settings. A series of confidential reports were prepared between November 2010 and March 2011 for consideration by the GMC and their technical advisers. This document outlines the methods used to assess the performance of the questionnaires and provides a summary of the key findings and recommendations arising from the main survey work.

2 Overview of Methods

2.1 GMC patient and colleague questionnaires

Both questionnaires include items relating to the GMC's core guidance on the principles and values to which it requires registered doctors to adhere.[8] The GMC Colleague Questionnaire (CQ) comprises 19 core items and has been designed to be administered online, although a paper version is available. The GMC Patient Questionnaire (PQ) comprises 9 core items and is designed to be administered as a post-consultation 'exit survey'. The content of the questionnaires is summarised in Appendix 1.

2.2 Survey

All non-training grade doctors from eleven sites in England and Wales were invited to take part in the survey between August 2008 and January 2011. The sites included primary care organisations, acute hospital and mental health trusts, and the independent sector.

Participating doctors nominated up to 20 colleagues (10 medical; 10 non-medical) who could provide structured feedback on their professional performance using the GMC CQ. In addition, if the doctor's role included regular patient contact, they distributed a GMC PQ to up to 45 consecutive patients. Doctors also completed a self-assessment questionnaire, which mapped onto the content of the PQ and CQ.

The procedure for collecting the feedback has been described elsewhere.[9] Client Focussed Evaluation Programme UK Surveys (CFEP-UK) coordinated the recruitment of doctors, and the collection and management of the survey data. Anonymised data sets were provided to the academic team for analysis.

Doctors who completed the MSF process received a personalised feedback report which summarised the results of their surveys. They were encouraged to identify a supporting medical colleague from the outset, and to discuss their results with them. Doctors could also choose to use the report within their formal appraisal process.

2.3 Data analysis

Using the large data sets obtained from the main survey, the psychometric properties of the PQ and CQ were investigated via a range of statistical techniques based on Classical Test Theory, Generalisability Theory and Rasch analysis. These included:

- Examining overall response rates, and core item completion rates to assess the acceptability of the questionnaires;
- Reviewing the frequency and distribution of responses;
- Calculating a Cronbach's alpha coefficient for each questionnaire to explore internal consistency;
- Conducting principal components analysis (factor analysis) to identify any sub-scales within the questionnaires;
- Completing a Generalisability (G) study to assess the sources of measurement error and determine the reliability of the questionnaires; and a Decision (D) study to estimate the reliability that might be obtained by varying the number of questionnaire items or the number of assessors;
- Undertaking Rasch analysis to test the measurement properties of the rating scales. Rasch represents a rigorous and sophisticated analytical approach likely to expose weaknesses in survey instruments.

A series of parallel studies and analyses explored additional aspects of the performance of the GMC questionnaires:

- To explore test-retest reliability, patients and colleagues of a sub-sample of doctors completed two questionnaires approximately two weeks apart. For the patient study, all PQs were completed via post; for the colleague study all CQs were completed online. Intraclass correlations were calculated to assess the temporal stability of patient and colleague ratings on the core questionnaire items (5-point scales) and for respondent-level PQ and CQ scores. Kappa statistics were calculated to assess the stability of responses on the binary scales.
- To explore convergent validity, patients and colleagues of a sub-sample of doctors completed extended questionnaires. For patients, the questionnaire booklet included items from the PQ, the Doctors' Interpersonal Skills Questionnaire (DISQ)[13] and the Patient Enablement Inventory (PEI),[14] which measure similar (DISQ) or related (PEI) concepts. For colleagues, the extended questionnaire included items from the CQ and the Colleague Feedback Evaluation Tool (CFET),[15]

which measure similar aspects of performance. Spearman's correlation coefficients (rho) were calculated to explore the strength of the relationship between mean scores achieved on the PQ, DISQ and PEI (patient study), and the relationship between mean scores achieved on the CQ and CFET (colleague study).

- To determine whether sampling biases might exist, (i) the effect of patient sample characteristics on core PQ item ratings and (ii) the effect of colleague sample characteristics on core CQ item ratings were explored using regression modelling.
- Multivariate modelling was undertaken of patient and colleague feedback, after deriving two summary scores for each doctor based on mean item scores. Modelling took account of the characteristics of the doctor, and of the patient or colleague samples, with a view to identifying independent predictors of the doctor's summary scores.

2.4 Additional sub-studies and sub-analyses

In parallel to the main survey work, a series of sub-studies and sub-analyses were conducted to examine specific questions relating to the implementation of MSF in applied settings. Whilst the findings of these sub-studies and analyses are not included in this Executive Summary, the focus of each study is summarised in Appendix 2.

3 Survey findings

3.1 Doctor participation and data collection

- By 17 January 2011, all eleven sites had successfully undertaken the GMC MSF work in a census sample of non-training grade doctors. The original aim was to secure the participation of 1000-1250 doctors.
- Doctor recruitment rates ranged from 30% to 65% across organisations. Participation rates varied between clinical specialties.
- Across all sites, 2454 doctors were eligible for inclusion and approached, of whom 1067 doctors (43%) agreed to take part and 1065 doctors contributed some questionnaire data.
- 1057 (99%) doctors provided colleague data, 922 provided patient data (87%) and 1037 provided self-assessment data (97%).
- Feedback was obtained from 30333 patients (using the PQ) and 17012 colleagues (using the CQ).
- 918/1065 (86%) doctors returned both PQ and CQ data, of which 777/918 (85%) provided sufficient returns on both surveys (≥22 PQs and ≥8 CQs) using sample targets set during our earlier work[9] to form a reliable estimate of the doctor's performance.
- A median time of 49 days (lower quartile (LQ) 29 days, upper quartile (UQ) 82 days) was required to return $\geq$ 22 PQs, and 8 days (LQ 4 days, UQ 14 days) were required to return $\geq$ 8 CQs.
- Our previous research suggested that general practitioners working in smaller practices were concerned about potential difficulties in recruiting sufficient colleagues and patients for the purpose of MSF. Our analysis identified that this did not appear to be the case; participant doctors from small practices recruited broadly similar numbers of patients and colleagues as doctors from larger practices.

3.2 Understanding the questionnaire performance

3.2.1 Acceptability and distribution of ratings

- The PQ and the CQ are generally acceptable to the patients or colleagues who complete them. There was minimal missing or spoilt data (<1% for PQ core items; 1% to 3% for CQ core items).
- Responses on the PQ and CQ were highly skewed towards favourable impressions of the doctor's performance, with the majority of respondents selecting the 'good' or 'very good' response options across the core PQ and CQ items. Less than satisfactory ratings of the doctor's performance were rare (<1% across PQ core items; ≤1% on CQ core items).
- The use of 'does not apply' as a response option varied from <1% to 11% across individual items in the patient questionnaire. The use of 'don't know' as a response option varied from 1% to 28% across individual items in the colleague questionnaire.
- The use of the 'don't know' response option appeared logical and supported the validity of the CQ. Colleagues with an administrative/managerial role used the 'don't know' option more frequently than colleagues with other roles. When this response option was adopted, it appeared that colleagues were unable (rather than unwilling) to provide ratings on individual aspects of performance.

3.2.2 Internal consistency and structure of the questionnaires

- Both questionnaires have good internal consistency with Cronbach's alphas of 0.865 for the PQ and 0.938 for the CQ.
- Both questionnaires appear to measure two broad domains of a doctor's professional performance which might be summarised as: (i) 'interpersonal, clinical and organisational skills'; and (ii) 'probity' (the latter also includes health from the CQ).

3.2.3 Test-retest reliability

- Using a classical approach to analysis, the GMC questionnaires appear to have acceptable test-retest reliability over a two-week period: responses on both questionnaires were stable over time.
- The intraclass correlation (ICC) relating to the overall (patient-level) PQ score was high (0.834; 95% confidence interval (CI): 0.792 to 0.868) indicating that patients'

responses on the PQ were very similar at both time points. ICCs for individual PQ items were satisfactory on items relating to the doctor's consulting skills (ranging from 0.627 to 0.732) but were slightly lower for those relating to their probity (ranging 0.582 to 0.629).

- The ICC relating to the overall (colleague-level) CQ score was high (0.851; 95% CI: 0.803 to 0.888), indicating that colleagues' responses on the CQ were very similar at both time points. ICCs for individual CQ items were satisfactory on items that related to the doctor's clinical, organisational, teaching and communication skills (range 0.602 to 0.768) but were lower on the probity/health items (ranging 0.450 to 0.596).
- Minimal variation over time in the use of the 'Does not apply' (PQ) and 'Don't know' (CQ) response options was observed and kappa statistics were satisfactory.

3.2.4 Convergent validity

- Using a classical approach to analysis, the GMC questionnaires appear to be a valid measure of doctors' performance when respondents' PQ/CQ ratings are compared with ratings on other measures designed to assess similar aspects of a doctor's performance.
- Patient respondents' overall mean scores on the PQ and on a second measure (DISQ) designed to assess similar aspects of a doctor's performance were strongly correlated (rho=0.629, $p<0.001$).[16]
- The weaker correlation[16] (rho=0.314, $p<0.001$) between overall mean scores on the PQ and the PEI suggest that the two instruments are measuring related, but distinct aspects of a doctor's performance. This finding was in line with our a priori hypothesis.
- Colleague respondents' overall mean scores on the CQ and on a second measure designed to measure similar aspects of a doctor's performance (CFET) were strongly correlated (rho=0.808, $p<0.001$).[16]

3.2.5 Generalisability (G) and Decision (D) studies

- Analyses based on Generalisability Theory and the associated Decision (D) studies indicated that reliability (G coefficient) in excess of 0.70 would be attained with the return of 34 PQs or 15 CQs.
- These revised sample sizes reflect the 'naturalistic' settings in which the data were obtained, the use of untrained raters/observers to provide data, and a more

conservative approach to the assessment of generalisability adopted in this study compared with our earlier work.[9]

3.2.6 Rasch Analysis

- This is the first occasion that Rasch analysis has been used to investigate the properties of questionnaires designed to capture information about a doctor's professional practice. As far as we are aware, no other questionnaires of similar intent have undergone such a rigorous process of examination.
- Analysis of the GMC questionnaires highlighted the potential utility of the PQ and the CQ in the assessment of doctors, but also highlighted the need for caution in respect of inferences made about a doctor's professional practice when drawing on data obtained using these instruments.
- Rasch analysis identified three areas of particular concern in respect of the PQ and CQ. These related to: (i) the targeting of scales to the samples of doctors taking part; (ii) problems with some of the response categories and response descriptors, most notably the 'less than satisfactory' response option; and (iii) cohesiveness of the construct being mapped by items within the instruments.

3.3 Modelling individual respondent feedback

3.3.1 Patient respondent characteristics

- Regression modelling was undertaken to explore the determinants of patient responses. Seven potential patient-related predictors of response to the nine core PQ items were investigated.
- After correcting for clustering of responses by doctor, two patient characteristics were the strongest independent predictors of *more favourable* patient responses: (i) reporting higher importance of the visit; and (ii) seeing their 'usual' doctor.
- The patient's age also predicted responses on some of the PQ items: older patients tended to have *more favourable* views of doctors performance when compared with younger patients.
- Two other characteristics predicted *less favourable* patient responses on some of the PQ items: (i) returning the survey via post; and (ii) the patient reporting their ethnicity as 'non-White'.
- Neither gender nor the type of respondent (the patient vs. another person responding on behalf of the patient) predicted ratings on any of the nine PQ items.

3.3.2 Colleague respondent characteristics

- Regression modelling was undertaken to explore the determinants of colleague responses. Seven potential colleague-related predictors of response to the 18 core CQ items were investigated.
- After correcting for clustering of responses by doctor, the professional role of the respondent was an independent predictor of responses on some of the core CQ items. Managers/administrative staff, and non-medically qualified healthcare professionals had *more favourable* views of doctors' performance than medical colleagues (whether fully trained or not).
- The frequency of the colleague's contact with the doctor was also a predictor of responses on some of the CQ items: less frequent contact was associated with *less favourable* colleague responses.
- Questionnaire return method was an independent predictor on only one CQ item, with colleagues who provided online feedback giving a more favourable response than those returning a paper questionnaire.
- The colleague respondent's age, gender, ethnic group, and recency of contact with the doctor were not independent predictors of their response to core CQ items.

3.4 Development of benchmarks

- Benchmark scores can be calculated for the core PQ and CQ items. As well as calculating 'generic' benchmarks (for all doctors, regardless of clinical role), there is now sufficient data to calculate benchmarks at the level of *setting* (i.e. primary vs. secondary care).
- For some (but not all) clinical specialties, benchmarks at a *specialty*-group level (general practice vs. medicine/paediatrics vs. surgery/obstetrics and gynaecology vs. psychiatry) have also been calculated.
- As more doctors complete the MSF process, it should be possible to calculate benchmarks for individual specialties. In the meantime, caution needs to be exercised in deriving such benchmarks, particularly where small numbers of doctors from a given specialty have contributed to the dataset.
- There is evidence of a range of performance across questionnaire items, setting and specialty groupings.

- There is a notable skew in the benchmark data, with only small differences in scores being evident between the lower quartile and upper quartile for all items on the PQ and the CQ.

3.4.1 Exploring variation in doctor performance

- Using a norm-referenced approach, of 532 doctors who had returned at least 35 patient questionnaires, 50 (9%) had at least one of the nine core PQ items in which the mean item score was a statistical 'outlier' at the lower end of the scoring range. All doctors who had statistical outliers for three or more of the nine (individual) core PQ items also had an overall patient summary score which was statistically outlying.
- Using a norm-referenced approach, of 804 doctors who had at least 15 colleague questionnaires returned, 150 (19%) had at least one of the 18 core CQ items in which their mean item score was a statistical 'outlier' at the lower end of the scoring range. All doctors who had statistical outliers for eight or more of the eighteen (individual) core CQ items also had an overall colleague summary score which was statistically outlying.
- Only one doctor was a statistical outlier at the lower end of the scoring range on both patient summary and colleague summary scores; 18 doctors were outliers on colleague scores only, and 14 doctors were outliers on patient scores only.

3.4.2 Exploring variation in doctors' patient-derived scores

- Seven doctor characteristics were associated with variation in the doctors' overall summary scores based on patient feedback. These included the doctor's age, gender and ethnic group, the region from which they obtained their primary medical qualification, their clinical specialty, current contractual role, and their locum status.
- Seven patient sample characteristics were significantly associated with the doctors' overall summary scores based on patient feedback. These included the proportions of patients who were female; were aged under 21 years; were aged over 60 years or more; regarded their visit as 'very important'; were seeing their 'usual' doctor; and the proportion of patients whose ethnic group was 'White', or whose ethnic group was 'Asian'.
- A regression model accounting for both patient sample characteristics and doctor characteristics identified that five variables were independent predictors of doctors' patient-derived summary scores on the PQ. These included the doctor's clinical specialty, the region from which their primary medical qualification was obtained,

and the proportions of patient respondents who were 'White', who regarded their visit as 'very important', or who reported they were seeing their 'usual' doctor.

3.4.3 Exploring variation in doctors' colleague-derived scores

- Seven doctor characteristics were associated with variation in the doctors' overall summary scores based on colleague feedback. These included the doctor's age, gender, ethnicity, clinical specialty group, current grade, locum status and the region in which their primary medical qualification was obtained.
- Eight colleague sample characteristics were associated with the doctors' overall summary scores based on colleague feedback. These included the proportions of colleagues who were aged under 30 years, who were aged 60 years or more, who currently worked with the doctor, who reported being in daily or weekly contact with the doctor during their period of working together, completed a paper version of the CQ, or the proportion of colleagues whose ethnic group was 'White', or whose ethnic group was 'Asian'.
- A regression model accounting for both colleague sample-derived variables and doctor characteristics identified that five variables were independent predictors of doctors' colleague-derived scores on the CQ. These included the doctor's clinical specialty group, the region in which their primary medical qualification was obtained, their contractual role, their status as a locum, and the proportion of colleague respondents who reported daily or weekly contact with the doctor.

4 Recommendations

The patient and colleague questionnaires were initially developed within the GMC. The domains mapped to the content of core guidance on doctors' professionalism – "Good Medical Practice".[8] This research investigated the potential utility of these questionnaires as a means by which doctors might gather evidence regarding their professional practice as assessed and reported by their patients and colleagues.

4.1 Patient assessments

The GMC patient questionnaire, whilst having a number of limitations, is a sufficiently robust instrument for its use in the preliminary rollout of obtaining patient feedback for the purposes of revalidation.

Detailed analysis based on Rasch modelling identified some concerns regarding the scale descriptors adopted and the degree to which the questionnaire assessed the theoretical concept of a doctor's professional practice and behaviours. Exploration of alternative forms of scale descriptors is necessary with a view to improving the scaling properties of the questionnaires. Further research is required to define the concept of 'professionalism' in doctors.

There was evidence of systematic bias in reporting amongst some groups of patient respondents. The composition of the patient sample may have implications for a doctor when attempting to interpret their data compared with benchmarks arising from normative samples of patient ratings.

Certain doctor characteristics appeared to predict systematic variation in patient assessments. This is an important observation – but it is unclear to what extent these observations reflect "true" variation in professional performance, or raise the possibility of systematic bias in patients' reports of doctor performance for other, non-performance-related reasons.

In interpreting data arising from such surveys of patients, consideration should be given to the possibility of systematic bias in a patient's report based on non-clinical aspects of care, as well as the socio-demographic profile of the patient sample.

4.2 Colleague assessments

While Rasch analysis identified some concerns regarding the scale descriptors used within the colleague questionnaire, the colleague questionnaire generally performed more robustly than the patient questionnaire.

The GMC colleague questionnaire is a sufficiently robust instrument for its use in the preliminary rollout of obtaining colleague feedback for the purposes of revalidation.

There was evidence of systematic bias amongst some groups of colleague respondents in respect of assessments provided.

Based on the profile of responses obtained, the guidance provided to colleagues regarding the identification and mix of colleagues who might participate in the MSF process appears appropriate.

Certain doctor characteristics appeared to predict systematic variation in colleague assessments. Again, it is unclear to what extent these observations reflect "true" variation in professional performance, or raise the possibility of systematic bias in colleagues' reports of doctor performance for other reasons.

In interpreting data arising from such surveys of colleagues, consideration should be given to the possibility of systematic bias in a colleague's report based on non-clinical aspects of care, as well as the extent of colleagues' familiarity with the doctor.

4.3 GMC tools

Both the GMC patient and colleague questionnaires represent instruments which would provide a reasonable basis for the collation of evidence regarding a doctor's professional performance.

In view of the highly skewed data, and importantly, the lack of availability of robust specialty-specific benchmarks, caution should be exercised in respect of actions which arise as a result of undertaking multi-source and patient feedback using these instruments.

Our recommendation is that these instruments should be used in a largely formative basis for the first two or three years whilst mandatory participation in MSF becomes the norm. We also recommend that the MSF process should include the completion of a self-assessment tool to aid self-reflection.

4.4 Reporting to doctors and actions arising

Guidance is necessary within the context of a report, particularly with regard to the assignation of doctors' ranked performance. Careful explanation is also required in respect of doctor scores on individual items, and in respect of any summary scores derived from items across the whole questionnaires.

Doctors are sensitive to the results obtained. We would advise that doctors should have a supporting medical colleague with whom they can discuss the results in an environment which is non-threatening, supportive and informative.

Both questionnaires and the data arising from them have strong formative potential. At this stage, the use of data arising from these questionnaires (or any other type of MSF questionnaire) for summative purposes needs to be approached with caution.

Initially, we advise that these questionnaires may provide the basis of a useful screening process by which doctors and/or their appraisers might be alerted to areas in which remedial action might be taken.

The questionnaires do have the potential to "rank" doctors' performance, and it is likely that only a small number of doctors at the lower end of the ranking scale should be considered for review of their performance in the context of appraisal and other collated evidence of performance.

However, we emphasise that, even when a doctor is identified as being statistically located at the lower end of the performance spectrum, it is important that the MSF results are interpreted within the context of the setting in which a doctor is practising, their personal characteristics, and the mix of patients for whom they are providing care.

The threshold of performance which might trigger any review of a doctor's performance within the context of appraisal is not absolute and will be determined following a consideration of the robustness of the processes established around multi-source and patient feedback, and in relation to the resource implications which might arise.

Further discussion is necessary with the relevant regulatory authorities in respect of standard setting and appropriate actions that might arise.

A reasonable approach might be to consider that doctors whose performance in any individual questionnaire item fell below two standard deviations of the mean score (for that item) might be advised of their specific performance and encouraged to take appropriate remedial/educational action.

Doctors whose performance was more markedly lower (perhaps falling below three or more standard deviations below the mean score for an item) might be strongly encouraged to undertake such action and advised to undertake further testing within a defined timeframe.

Promoting the formative potential and the links to guided continuing professional development provides an important potential opportunity for the positive use of information arising from the questionnaires.

4.5 Rollout

Clinical care is provided in a very wide range of settings. Some groups of doctors face particular challenges in undertaking and completing an MSF process.

It is important to be alert to the specific challenges some doctors will face in undertaking MSF on account of their clinical environment and to provide the necessary support for such doctors in undertaking MSF.

Were these questionnaires to be considered for use on a large scale, it is vital that an external organisation with sufficient experience of handling confidential data is employed to provide centralised support to doctors and organisations in which data collection is ongoing.

The administrative workload involved in undertaking these surveys is significant. There is a need for both central co-ordinated support delivered by a professional survey organisation and local administrative support provided to doctors undertaking these surveys.

Our experience of the online colleague survey suggests it was both efficient and significantly non-intrusive to be acceptable to potential respondents. However, appropriate arrangements are necessary to support individuals – perhaps as many as 20% of potential colleague respondents – who wish to use paper-based return of the colleague questionnaire.

Security of data is paramount and appropriate arrangements are required to ensure that doctors do not have access to the questionnaires completed by their patients or colleagues.

4.6 Benchmarks and standard setting

The development of robust benchmarks is vitally important to allow the comparison of performance amongst doctors from different geographical and clinical contexts.

Given our observations on the potential for respondent bias according to socio-demographic variables, suitably robust benchmarking of scores may be necessary for doctors providing clinical care in settings of varying socio-demographic mix.

Achieving robust benchmarking will only be possible by the use of one, or a very limited number of questionnaires.

We would advise the adoption of a limited number of patient and colleague questionnaires whose performance characteristics are well understood.

Whilst recognising some limitations, we would advise that the GMC patient and colleague questionnaires are suitable instruments, at least in the first instance.

Should a range of MSF instruments be endorsed, rigorous evidence regarding their reliability and validity should be mandated to ensure that doctors, society, and the regulatory authorities can be confident in their results, and also appreciate the uncertainties which are inherent in any similar system. That evidence must be subjected to the normal processes of peer review, and publication in the scientific literature.

4.7 Further research

It is clear that questionnaires targeting patients and colleagues have potential in informing the evidence base around a doctor's professional status and fitness to practise.

Development of questionnaires and supporting MSF survey processes which have validity, reliability, educational impact, acceptability, feasibility and a reasonable cost is of importance.

This research informs a number of those elements but it is vital to recognise that attaining questionnaires which have the confidence of patients, wider society, the medical profession and professional regulators is a process, not an event.

Research is required to examine the cost and benefits to doctors and society in undertaking these processes.

Further research exploring, refining and informing the use and potential of multi-source feedback from the patients and colleagues of doctors is therefore essential.

5 References

1. General Medical Council. Revalidation: *The way ahead - Consultation Document.* London: General Medical Council, 2010

2. General Medical Council. Revalidation: *The way ahead - Response to our revalidation consultation.* London: General Medical Council, 2011.

3. General Medical Council. *The Good Medical Practice Framework for appraisal and revalidation.* London: General Medical Council, 2011.

4. General Medical Council. *Supporting information for appraisal and revalidation.* London*:* General Medical Council, 2011.

5. General Medical Council. Revalidation: *The way ahead. Annex 3 - GMC principles, criteria and key indicators for colleague and patient questionnaires in revalidation.* London: General Medical Council, 2010.

6. Chisholm A, Askham K, 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.

7. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004; **328**(7450): 1240-1245.

8. General Medical Council. *Good medical practice: Guidance for doctors*. London: General Medical Council, 2006.

9. Campbell JL, Richards SH, Dickens A, et al. Assessing the professional performance of UK doctors: An evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care* 2008; **17**: 187-193.

10. Kilminster S, Pell G, Roberts T. *Patient and colleague questionnaires: Validation report to the GMC*. Leeds: University of Leeds, Medical Education Unit, 2005.

11. MORI Social Research Institute. R*evalidation questionnaires testing: Qualitative research findings*. London: General Medical Council, 2004.

12. Lockyer J, Fidler H. *Comparison of colleague and patient multisource feedback instruments designed for GPs in UK.* London: Royal College of General Practitioners, 2010.

13. Greco M, Cavanagh M, Brownlea A, McGovern J. The Doctors' Interpersonal Skills Questionnaire (DISQ): A validated instrument for use in GP training. *Education for General Practice* 1999; **10**: 256-264.

14. Howie JGR, Heaney DJ, Maxwell M. *Measuring quality in general practice: Pilot study of a needs process and outcome measure. Occasional Paper.* London: Royal College of General Practitioners, 1997.

15. Campbell J, Narayanan A, Burford B, Greco M. Validation of a multi-source feedback tool for use in general practice. *Educ Prim Care* 2010;**21**:165-79.

16. Cohen J. A power primer. *Psychol Bull* 1992; **112**: 155-159.

17. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof* 2003; **23**: 4-12.

18. Sargeant, J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: Perceptions of credibility and usefulness. *Med Educ* 2005, **39**: 497-504.

19. Sargeant, J, Mann K, Sinclair D, et al. Challenges in multisource feedback: Intended and unintended outcomes. *Med Educ* 2007; **41**: 583-591.

Appendix 1: Summary of content of GMC questionnaires

(i) Content of the GMC Patient Questionnaire (PQ)

Item	PQ items: Instructions and item stems
	Contextual and descriptive items
Q1	Who is filling in the patient questionnaire
Q2	Reason(s) why the patient saw the doctor
Q3	How important the visit was to patient's health/wellbeing
Q8	Was the patient's visit with their usual doctor
Q10-12	Sex, gender, ethnicity
	Core performance evaluation items
	Please rate your doctor on the following areas: [a]
4a	Being polite
4b	Making you feel at ease in his / her presence
4c	Listening to you
4d	Assessing your condition
4e	Explaining your condition and treatment
4f	Involving you in decision about your treatment
4g	Providing or arranging treatment for your
	Please decide how far you agree with the following statements: [b]
5a	This doctor will keep information about me confidential
5b	This doctor is honest and trustworthy
	Summative items [c]
Q6	I am confident about this doctor's ability to provide care
Q7	I would be completely happy to see this doctor again
	Free text
Q9	*Please feel free to add any other comments you have about this doctor*

[a] Valid evaluative response categories (scale number): 'Poor' (1), 'Less than satisfactory' (2), 'Satisfactory' (3), 'Good' (4), 'Very good' (5).
[b] Valid agreement response categories (scale number): 'Strongly disagree' (1), 'Disagree' (2), 'Neutral' (3), 'Agree' (4), 'Strongly Agree' (5).
[c] Binary response categories: 'Yes' (1), 'No' (2)

(ii) Content of the GMC Colleague Questionnaire (CQ)

Item	CQ items: Instructions and item stems
	Core performance evaluation items
	Please rate your colleague in each of the following areas: [a]
Q1	Clinical knowledge
Q2	Diagnosis
Q3	Clinical decision making
Q4	Treatment (including practical procedures)
Q5	Prescribing
Q6	Medical record keeping
Q7	Recognising and working within limitations
Q8	Keeping knowledge and skills up to date
Q9	Reviewing and reflecting on own performance
Q10	Teaching (students, trainees, others)
Q11	Supervising colleagues
Q12	Commitment to care and wellbeing of patients
Q13	Communication with patients and relatives
Q14	Working effectively with colleagues
Q15	Effective time management
	Please decide how far you agree with the following statements: [b]
Q16	This doctor respects patient confidentiality
Q17	This doctor is honest and trustworthy
Q18	This doctor's performance is not impaired by ill health
	Summative item [c]
Q19	This doctor is fit to practise medicine
	Free text
Q20	*Please add any other comments you want to make about this doctor.*
	Demographic and contextual items
Q21	Gender
Q22	Age group
Q23	Professional role
Q24	How recently familiar with the doctor's clinical practice
Q25	Frequency of contact with the doctor
Q26	Ethnic group

[a] Valid evaluative response categories (scale number): 'Poor' (1); 'Less than satisfactory' (2); 'Satisfactory' (3); 'Good' (4); 'Very good' (5).
[b] Valid agreement response categories (scale number): 'Strongly disagree' (1); 'Disagree' (2); 'Neutral' (3); 'Agree' (4); 'Strongly agree' (5).
[c] Binary response categories: 'Yes'' (1); 'No' (2).

Appendix 2: Overview of other sub-studies and analyses

In parallel to the main survey work, a series of sub-studies were conducted to explore specific issues relating to the implementation of MSF in applied settings. In addition, further sub-analyses were conducted using the main survey data sets, to explore the level of insight that doctors have into their own performance.

Whilst the findings of these sub-studies and sub-analyses are not included in this Executive Summary, the focus of each is outlined below. In future, we aim to seek publication of the data derived from these sub-studies and analyses in relevant peer-reviewed academic journals.

Comparison of doctors' self-assessment ratings with those provided by patients and colleagues

Previous research has suggested that the extent to which an individual's assessment of their own performance matches that obtained from other sources may affect the likelihood that they will take action on the feedback.[17-19] In the light of this observation, a sub-analysis was conducted using the main survey data sets, to explore the relationships (levels of agreement) between doctors' self-ratings and the ratings provided by their patients and colleagues on the core PQ/CQ items. The effect of a range of doctor characteristics on any observed differences between the perceptions of doctors and those of their patient/colleague assessors were investigated using regression analyses.

Development of personalised feedback reports for doctors

This action research study sought to develop the personalised reports that are used to feed back the results of patient and colleague surveys to doctors. Across two study cycles, the views and experiences of a sub-sample of doctors who had completed the GMC MSF process were obtained. The themes arising from the telephone interviews were used to guide the further development and refinement of the feedback report template.

Views and experiences of the MSF process as part of appraisal

This qualitative study used telephone interviews to explore the experiences of a sub-sample of doctors' and appraisers who had used the GMC questionnaires in the appraisal process. Their views of the future use of conducting MSF for appraisal and revalidation were also explored.

Feasibility and acceptability of the MSF process for anaesthetists

As part of the main survey, non-training grade anaesthetists from six participating organisations were invited to pilot the GMC MSF process. Quantitative and qualitative data was collected to explore the feasibility and acceptability of the MSF process in this clinical setting, and to identify specific challenges that anaesthetists might encounter in attempting to obtain reliable feedback from patients and colleagues.

Views and experiences of the support mechanisms available

An online survey of a sub-sample of participating doctors and their nominated supporting medical colleagues was initiated in November 2010. It sought to identify the type and level of support received by the doctors at the end of the MSF process, and to seek views on the value of such support.