

Feasibility of developing and selecting criteria for the assessment of clinical performance

NEIL JOHNSON

SUMMARY. *This study was undertaken to assess the feasibility of one method of initiating the setting of standards in United Kingdom general practice by the development and selection of criteria for clinical care. Three clinical topics were chosen. An 'academic phase' was used to develop precise criteria, which formed the basis of a draft questionnaire for each of the three topics. A 'consultation phase' followed in which the three questionnaires were sent to a randomly selected sample of general practitioners throughout the UK. Although 53% of the 62 general practitioners invited were willing to be involved in the academic phase only 13% were able to attend a meeting. Of the 262 general practitioners sent one of the questionnaires 63% responded. Using a strict definition of consensus one quarter of the criteria were identified by the respondents to the questionnaire as important or not important; using a less strict definition a further one quarter to one half of the criteria were identified.*

The study demonstrates that it is possible to develop and select criteria in this way. If this approach were to be developed for performance review, criteria would need to be selected, appropriate target levels of performance set, and suitable methods for their application identified.

Keywords: *performance review; professional competence; clinical knowledge; clinical experience; protocols.*

Introduction

THE drive for improving quality within general practice in the United Kingdom has come both from inside¹ and outside² the profession. Within general practice the drive has been particularly directed towards reaccreditation³ and the granting of fellowship of the Royal College of General Practitioners by assessment.⁴ Part of the drive for improving quality in medicine has seen a rapidly increasing emphasis on audit, although alternative approaches are under consideration.⁵

A fundamental component of audit is the setting of standards, defined as benchmarks against which performance can be measured.⁶ However, in a review of studies examining the performance of general practitioners Rethans and colleagues noted that in some reports the standards were not made explicit; in others although the standards were explicit they had been created by specialists or academic general practitioners and thus may not have been appropriate to everyday general practice.⁷

One alternative is to invite general practitioners themselves to set the standards. Standards have been set in this way in both the Netherlands⁸ and the UK.^{9,10} However, in both of these initiatives the general practitioners involved were selected for particular attributes — in the Netherlands they had a special interest in the subject under consideration⁸ while in the UK study all were trainers.^{9,10} In the study described here general practitioners were

selected at random in an attempt to encompass the majority of the views present in general practice. Such an approach has recently been supported by some of those involved in the Dutch initiative.¹¹

One way of simplifying the process of standard setting is to divide the process into two.¹² First, the selection of clinical behaviours which are markers of high quality care (criteria). Secondly, the determination of the levels to which these criteria should be performed. This paper reports a formal assessment of the feasibility of one method of selecting criteria for the assessment of clinical performance.

Method

Three clinical topics were chosen to reflect common clinical situations, to apply to patients of different ages and both sexes, to cover both proactive and reactive care, and to encompass acute and chronic disease. They were acute earache in a child aged under 10 years, long-term hypertensive care, and oral contraceptive care.

Criteria development and selection

Academic phase. General practitioners who had published work in the relevant subjects during the period 1985–89 were identified by a search of the National Library of Medicine's database, MedLine. In March 1990 these doctors were invited to a meeting together with general practitioners from academic departments of general practice in Wales and the south west and Midlands areas of England (the latter group were asked to indicate their areas of publication and interest in their reply). Those who could attend met in small groups, one for each subject, each chaired by N J — two of the meetings were held in Birmingham, the third in Southampton. Each subject was considered for two half days. Participants were invited to identify aspects of care which might be used as criteria and to develop a precise definition for each potential criterion. Those who could not attend were asked to comment by post on the criteria developed by the small groups.

The criteria developed formed the basis of a draft questionnaire for each of the three topics. The ordering of the questions on the questionnaire was not influenced by the views of the small groups on the importance of the criteria. Each questionnaire was piloted to 10 general practitioners (local colleagues of N J) and using their comments to modify only the clinical vignettes and the layout of the questionnaire but not the criteria a final questionnaire was drawn up. None of the potential criteria developed by the small groups was excluded from the final questionnaire.

Consultation phase. Three hundred general practitioners from throughout the UK were randomly selected from *The medical directory* for 1991.¹³ Seven were excluded, either because they had been involved in the original small group discussions or because they were known to have left general practice. In July to September 1991 the questionnaire on acute earache in a child aged under 10 years was sent to 97 doctors, that on long-term hypertensive care to 96 doctors, and that on oral contraceptive care to 100 doctors. A covering letter from N J was sent to each doctor with the questionnaire. A maximum of two reminders were sent at monthly intervals to non-respondents after cross-checking their address in *The medical register* for 1991.¹⁴

Each questionnaire began with a short summary of the clinical problem. Respondents were asked to indicate how important they

N Johnson, MA, MRCP, MRCPGP, RCGP research training fellow, Department of Public Health and Primary Care, University of Oxford. Submitted: 17 December 1992; accepted: 17 March 1993.

© *British Journal of General Practice*, 1993, 43, 499–502.

believed each criterion to be using a five-point scale. For example, part of the questionnaire about acute earache was as follows:

You see a child, aged under 10 years, who presents with earache. It is the first time that you have met the child or the parents.

I believe that it is important: to obtain a past history about ear problems

agree					disagree
1	2	3	4	5	

No definition of 'importance' was given. The questionnaire also sought information on the demographic characteristics of respondents and their practices. Each questionnaire comprised 70–90 questions, including two questions designed to distract respondents, for example the importance of examining the nails when seeing a child with earache.

Analysis of process

Participation. For the academic phase the willingness of those approached to participate and their attendance at meetings were formally recorded and used as measures of participation. For the consultation phase overall response rates were recorded and differences between respondents and non-respondents assessed in terms of sex and age. Respondents were also compared with national averages for general practitioners in the UK in terms of mean age and list size, and proportion of men/women, based on Department of Health data.¹⁵

Achievement of consensus. For the academic phase a subjective assessment of the degree of consensus among delegates was made. For the consultation phase an objective measure of consensus was made. The definition of consensus in the consultation phase had to reflect agreement that a potential criterion was either important or not important; only those criteria for which there were high levels of agreement would then be used for performance review. The definition used was that the median response was at one extreme of the scale (that is one or five) and that the difference between the 10th and 90th percentiles was one scale point or less. The sensitivity of this definition was tested by applying a less strict definition of consensus to see which additional criteria were then included. This less strict definition was that the median score was one, two, four or five (that is, not three) and that the range between the 10th and 90th percentiles was two scale points or less.

The results were analysed using the *SPSS-X* programme. The statistical significance of differences between proportions was based on the chi square test.

Results

Participation

Academic phase. Sixty two general practitioners were invited to take part in the academic phase. Of these, 33 (53%) were willing to be involved in the development of criteria (eight for acute earache in a child aged under 10 years, 12 for long-term hypertensive care and 13 for oral contraceptive care). Eight (13% of 62, 24% of 33) were able to attend the meetings (three for earache, three for hypertensive care and two for contraceptive care). There was no significant difference between attendance rates for the two groups of doctors approached — of the 26 general practitioners from university departments three attended (12%) while of the 36 identified from the literature five attended (14%).

Consultation phase. Of the 293 questionnaires sent to general practitioners, 31 were returned indicating the doctor had retired

from general practice or died; 164 of the remaining 262 doctors responded (62.6%). There was no difference between the response rates for the individual topics — 62% (55/89) for earache, 59% (50/85) for hypertension and 67% (59/88) for oral contraception. There were no significant differences between respondents and non-respondents in terms of sex or age for any of the three questionnaires.

There were no significant differences between respondents and national averages¹⁵ in terms of sex or mean list size. Respondents to the earache and contraception questionnaires were significantly younger than the national mean of 44.9 years (mean of 41.7 years ($P<0.05$) and 41.4 years ($P<0.01$), respectively) but the difference for the hypertension questionnaire was not significant.

Achievement of consensus

Academic phase. Subjectively, the consensus meetings appeared to allow discussion of a large number of potential criteria. Because the meetings were not being used to select criteria, the limited time (seven hours for each subject) was spent defining the criteria precisely; little difficulty was experienced in achieving this. It was recognized by all three groups that the actual selection of criteria would have been much more difficult and was best not undertaken in small groups (principally because of the peer pressure involved in small groups).

Consultation phase. The proportion of criteria falling into categories based on the level of consensus among general practitioners is shown in Table 1 for each subject. The criteria are divided into those of history taking, examination or investigation and management. The questions designed to distract are excluded from Table 1; for all six such questions there was either no agreement on their importance or they were agreed to be not important when the less strict definition was used.

When the strict definition of consensus was used between 24% and 27% of the potential criteria were considered important or not important. When the less strict definition was used a further 23–54% of potential criteria were included. There was no agreement about 22–52% of potential criteria. There was little agreement on the history criteria for earache or the examination/investigation criteria for oral contraceptive care.

Criteria selected

The criteria for earache agreed to be important using the strict definition were: to obtain a past history about ear problems; to make an assessment of 'how ill' the child is; to perform otoscopy to both ears; to examine the lymph nodes of the head and neck; to advise about oral pain relief; to avoid prescribing tetracyclines; to plan follow up if there is marked hearing loss; to arrange removal of a foreign body; to arrange follow up for barotrauma; and to refer urgently if mastoiditis suspected. The criteria for earache agreed not important using the strict definition were: to perform a full blood count; to undertake audiometry at presentation; to prescribe the same antibiotic again if otitis media fails to resolve after an initial course of that antibiotic; to undertake a McCormick toy test; and to prescribe oral steroids for glue ear.

The criteria for the first appointment for long-term hypertensive care agreed to be important using the strict definition were: to take at least three blood pressure readings before initiating therapy; to use a cuff bladder that encircles at least two thirds of the arm circumference; to ask if the patient smokes; to ask about a family history of coronary heart disease in family members under 60 years of age; to ask about a personal history of angina; to advise the patient to stop smoking (if the patient smokes); and to agree a target alcohol intake if the patient drinks excessive amounts of alcohol. The criteria for follow-up appointments

Table 1. Level of consensus for criteria considered.

	Total no. of criteria	No. (%) of criteria by strict definition		No. (%) of additional criteria by less strict definition		No. (%) of criteria about which no agreement ^a
		Important	Not important	Important	Not important	
<i>Earache</i>						
History	14	1 (7)	0	2 (14)	0	11 (79)
Examination/ investigation	16	3 (19)	3 (19)	0	2 (13)	8 (50)
Management	30	6 (20)	2 (7)	6 (20)	4 (13)	12 (40)
Total	60	10 (17)	5 (8)	8 (13)	6 (10)	31 (52)
<i>Hypertension</i>						
History	18	6 (33)	0	10 (56)	0	2 (11)
Examination/ investigation	23	3 (13)	0	12 (52)	0	8 (35)
Management	5	2 (40)	0	3 (60)	0	0
Total	46	11 (24)	0	25 (54)	0	10 (22)
<i>Oral contraception</i>						
History	20	6 (30)	0	9 (45)	0	5 (25)
Examination/ investigation	13	2 (15)	0	1 (8)	0	10 (77)
Management	11	4 (36)	0	4 (36)	0	3 (27)
Total	44	12 (27)	0	14 (32)	0	18 (41)

^aBy less strict definition.

agreed to be important were: to review the general well being of the patient; to check the smoking history; to ask if the treatment is causing any problems; and to check the blood pressure.

The criteria for the first appointment for oral contraceptive care agreed to be important using the strict definition were: to establish that the patient wants the oral contraceptive; to establish when the last menstrual period occurred; to check for a history of venous thrombosis; to check the smoking history; to check the blood pressure; to agree an appropriate starting day; to discuss other precautions to be used depending on the starting day chosen; to prescribe a pill with 35 µg oestrogen or less (unless an alternative is indicated clinically); and to discuss the routine of how the pill is taken. The criteria for follow-up appointments agreed important were: to ask about any problems; to check if the patient is still smoking; and to check the blood pressure.

Criteria for which there was consensus that they were important using the less strict definition of consensus included having a long-term management plan for glue ear for children with earache, checking the levels of urea and electrolytes and examining the chest at a first consultation for hypertensive care, and checking for a history of migraine at a first appointment for oral contraceptive care.

Discussion

There is now evidence that the setting of standards can improve both the process⁹ and outcome¹⁰ of general practice care. The aim of this study was to assess the feasibility of one method of selecting criteria which could then be used as one component of standards for general practice. The method chosen has been shown to be feasible but are the criteria selected reliable and valid, do they have any use, and what are the advantages and disadvantages of this method over alternatives?

Although no formal assessment was made of the validity of the criteria selected, they do have face validity, being consistent with published guidelines.¹⁶⁻²⁰ Furthermore, the finding that there was no agreement on the importance of most of the history criteria for earache and examination/investigation criteria for oral contraceptive care is consistent with the reality of everyday gen-

eral practice and supports the face validity of the criteria. If this method were to be more widely adopted it would be necessary to undertake formal tests of validity of the criteria chosen.

To those working outside primary care some of the criteria for which there was no agreement might appear surprising, such as, asking about a history of arterial disease when prescribing oral contraception. It would appear that primary care doctors need to be selective about what they do within their consultation, particularly because of time constraints.²¹ They exclude items that might be included by specialists and they also exclude items the relevance of which is unclear, such as breast and pelvic examination in contraceptive care. This selectivity emphasizes the need for criteria developed by general practitioners themselves rather than supplied by those who are not familiar with the constraints of everyday practice.

The response rate of 63% to the questionnaires used in the consultation phase is consistent with other studies using similar methods.^{11,22} The respondents were representative of all general practitioners in the UK in terms of sex and mean list size but the respondents to two of the questionnaires were slightly younger than the national average. It seems unlikely, however, that this small age difference had a major effect on the representativeness of the results.

The principal use of such criteria would be to allow general practitioners to assess their own performance using tools selected by their peers. In combination with target levels of performance (which could be set in a similar way) doctors could measure their own performance by reviewing notes, or analysing audio- or video-taped consultations. This would be more appropriate than attempts to measure competence.^{7,23} The criteria could also be used by external bodies (peer groups or non-medical bodies) in the external assessment of doctors' performance.⁴

The selected criteria could not be used as guidelines for care as guidelines need to be comprehensive rather than selective.²⁴ However, there is no reason why guidelines could not be selected in a similar way, using a strict definition of consensus to select crucial activities, and a less strict definition to select those aspects of management where greater clinical freedom is acceptable.

What are the advantages and disadvantages of the method described? The group discussion of the academic phase reduces individual bias; allows the development of criteria founded on sound scientific evidence; allows criteria without academic or practical significance to be excluded at an early stage; and allows refinement of the criteria. However, the practical difficulties of getting doctors to meet together meant that only one in eight of those approached actually attended. This phase is also expensive (travel/time approximately £300.00 per participant, venue approximately £200.00 per subject), and the time allotted (seven hours) was insufficient to allow both a systematic analysis of the literature and a structured consensus discussion. These practical disadvantages might be overcome by using the Delphi technique,²⁵ but the risk of this approach is that the chairperson can exert a greater influence.

The principal advantage of the consultation phase is that the criteria developed should reflect the views of the peers of those whose performance might be assessed. Unlike similar initiatives⁸⁻¹⁰ it avoids specific selection of respondents. It also avoids the peer pressure that exists when small groups attempt to achieve consensus and allows for the application of an objective measure of consensus. It is relatively cheap (approximately £4.00 per respondent) and data analysis is quick (approximately three minutes per respondent). It could also offer respondents the opportunity to provide comments to be fed back to the academic general practitioners to improve the criteria.

There are problems with using a consultation phase. Because of the method of selection the criteria developed may be seen to represent an empirical rather than normative standard. In addition, the development of criteria in this way prevents individual doctors from 'owning' the criteria. Ownership of standards seems to be vital to the improvement of performance and outcome.^{9,10} These disadvantages might be minimized if a range of target levels of performance were developed. Doctors could then select a target level of performance appropriate for their own practice. If this approach were to be more widely adopted it would be important to verify that the levels of performance adopted formed appropriate standards when combined with the consensus criteria. Finally some components of this phase are time consuming: in particular, questionnaire design and piloting, interpretation of results, and completion of questionnaires.

The other major limitation of the approach adopted here is that the criteria tested were frequently developed and chosen on the basis of a 'consensus hunch' rather than on good scientific evidence. Doubts have been expressed about the value of consensus as this may simply represent a drift towards the lowest common denominator.²⁶ The use of consensus criteria in general practice is therefore an area that requires long-term evaluation to determine whether positive outcomes are achieved. Unfortunately, much of everyday general practice is not supported by scientific evidence and it may be that, at present, a 'consensus hunch' is the best that can be offered for the development of criteria; it is almost certainly more useful than an 'individual hunch'.²⁷

If general practice of consistently high quality is to be achieved, the performance of general practitioners will continue to need to be assessed. General practitioners need to be encouraged to undertake assessment, using tools which they see as credible. The approach described here offers a feasible way of supplying such tools. If this approach were to be adopted it would need a suitable framework for developing criteria for all aspects of care (structure, process and outcome), the research to ensure that the criteria were valid, reliable and acceptable, a suitable method to ensure the setting of appropriate target levels of performance, and support (from inside and outside the profession) for its use.

References

1. Royal College of General Practitioners. *Quality in general practice. Policy statement 2*. London: RCGP, 1985.
2. Department of Health. *Terms of service for doctors in general practice*. London: DoH, 1989.
3. Gray DP. Reaccrediting general practice [editorial]. *BMJ* 1992; **305**: 488-489.
4. Royal College of General Practitioners. *Fellowship by assessment. Occasional paper 50*. London: RCGP, 1990.
5. Berwick DM, Enthoven A, Bunker JP. Quality management in the NHS: the doctor's role — I. *BMJ* 1992; **304**: 235-239.
6. *Medical audit tools*. Newcastle upon Tyne: Centre for Health Services Research, 1991.
7. Rethans JJ, van Leeuwen Y, Drop R, et al. Competence and performance: two different concepts in the assessment of quality of medical care. *Fam Pract* 1990; **7**: 168-174.
8. Grol RPTM, Mesker PJR, Schellevis FG (eds). *Peer review in general practice: methods, standards, protocols*. Nijmegen, Netherlands: Nijmegen University Department of General Practice, 1988.
9. North of England study of standards and performance in general practice. Medical audit in general practice. I: Effects on doctors' clinical behaviour for common childhood conditions. *BMJ* 1992; **304**: 1480-1484.
10. North of England study of standards and performance in general practice. Medical audit in general practice. II: Effects on health of patients with common childhood conditions. *BMJ* 1992; **304**: 1484-1488.
11. Rutten G, van de Laan J. Hypercholesterolaemia: setting a Dutch national standard. *Br J Gen Pract* 1992; **42**: 411-414.
12. Derry J, Lawrence M, Griew K, et al. Auditing audits: the method of Oxfordshire medical audit advisory group. *BMJ* 1991; **303**: 1247-1249.
13. *The medical directory*. London: Longman, 1991.
14. General Medical Council. *The medical register*. London: GMC, 1991.
15. Department of Health. *Statistics for general medical practitioners in England and Wales: 1979-1989. Statistical bulletin 4(4)91*. London: DoH, 1991.
16. Wilmot JF. Otitis media in general practice. *Practitioner* 1987; **231**: 1101-1107.
17. *British national formulary*. London: British Medical Association and Royal Pharmaceutical Society of Great Britain, 1992.
18. Treating mild hypertension. Report of the British Hypertension Society working party. *BMJ* 1989; **298**: 694-698.
19. Medical Research Council working party. MRC trial of treatment of mild hypertension: principal results. *BMJ* 1985; **291**: 97-104.
20. Guillebaud J. *Contraception: your questions answered*. London: Pitman, 1985.
21. Morell DC, Evans ME, Morris RW, Roland MO. The 'five minute' consultation: effect of time constraint on clinical content and patient satisfaction. *BMJ* 1986; **292**: 871-873.
22. de Melker RA, Kuyvenhoven MM. Management of upper respiratory tract infection in Dutch general practice. *Br J Gen Pract* 1991; **41**: 504-507.
23. Rethans JJ, Sturmans F, Drop R, et al. Does competence of physicians predict their performance? *BMJ* 1991; **303**: 1377-1388.
24. Schofield T, Hasler J, Barnes G. Implications for practice. In: Hasler J, Schofield T (eds). *Continuing care: the management of chronic disease*. Oxford University Press, 1990.
25. Linstone HA, Turoff M (eds). *The Delphi method. Techniques and applications*. Reading, MA: Addison Wesley, 1975.
26. Farmer A. Setting up consensus standards for the care of patients in general practice [editorial]. *Br J Gen Pract* 1991; **41**: 135-136.
27. Hammond KR. Towards increasing competence of thought in public policy formulation. In: Hammond KR (ed). *Judgement of decision in public policy formulation*. Boulder, CO: Westview Press, 1978.

Acknowledgements

This project was funded by a research training fellowship from the RCGP. I am also grateful to the Claire Wand Fund for substantial support. The guidance and advice of Drs David Mant, Ala Szczepura, Theo Schofield and John Wilmot were invaluable. I am most grateful to all the doctors who took part in both the academic and consultation phases.

Address for correspondence

Dr N Johnson, The Medical Centre, Badgers Crescent, Shipston-on-Stour, Warwickshire CV36 4BQ.