ORIGINAL ARTICLE

An experimental study of determinants of the extent of disagreement within clinical guideline development groups

A Hutchings, R Raine, C Sanderson, N Black

Qual Saf Health Care 2005;**14**:240–245. doi: 10.1136/qshc.2004.013227

See end of article for authors' affiliations

Correspondence to: Mr A Hutchings, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; andrew.hutchings@lshtm.ac.uk

Accepted for publication 10 April 2005

Objective: To assess the effect of design features and clinical and social cues on the extent of disagreement among participants in a formal consensus development process.

Methods: Factorial design involving 16 groups consisting of 135 general practitioners (GPs) and 42 mental health professionals from England. The groups rated the appropriateness of four mental health interventions for three conditions (chronic back pain, irritable bowel syndrome, and chronic fatigue syndrome) in the context of various clinical and social cues. The groups differed in three design features: provision of a systematic literature review (versus not provided), group composition (mixed versus GP only), and assumptions about the healthcare resources available (realistic versus idealistic). Disagreement was measured using the mean absolute deviation from a group's median rating for a scenario.

Results: None of the design features significantly affected the extent of disagreement within groups (all $p>0.3$). Disagreement did differ between treatments (closer consensus for cognitive behavioural therapy and behavioural therapy than for brief psychodynamic intervention therapy and antidepressants) and cues (closer consensus for depressed patients and patients willing to try any treatment).

Conclusion: In terms of the extent of disagreement in the groups in this study, formal consensus development was a robust technique in that the results were not dependent on the way it was conducted.

Guidelines for promoting good clinical practice are often developed using formal consensus methods.[1 2] The recommendations produced represent the views of groups about the research evidence and clinical opinion.[3 4] By definition, these methods seek to identify whether there is consensus or whether individual views diverge to such an extent that no recommendation can be made. While there have been a number of studies of the factors that may affect a group's view[5–9] (usually represented by the group mean or median), the factors that may affect the extent of disagreement (the spread of the distribution of ratings) have had relatively little attention.

The distinctive features of formal consensus methods are their transparency and structure, anonymity of individual appropriateness ratings, and feedback to panel members of preliminary results before final ratings are made.[5] The nominal group technique involves a ''nominal'' sample of about 10 people and includes a meeting to discuss areas of disagreement. An alternative is the mail only Delphi survey in which participants complete two or more rounds of questionnaires. The most widely used approach is a hybrid method developed by RAND[10] which involves a group of experts who first express their views independently via a mailed questionnaire. They then meet for a structured group discussion at which they clarify and explore areas of disagreement before completing the questionnaire again privately. There is no attempt to force the group to reach a consensus. The RAND method involves various definitions of ''agreement'' and ''disagreement'' based on the distribution of individual ratings. Where there is no ''disagreement'' (but not necessarily ''agreement''), the measure of central tendency (usually the median) indicates whether the group considers a treatment appropriate, inappropriate, or equivocal.

Previous research suggests that the extent of consensus or agreement depends on the group composition[11] and on the level of resources in the healthcare system.[5 12–15] The impact of providing the group with a systematic literature review has not been evaluated.[16]

Our aim was to investigate further some of the factors which may influence the extent of disagreement produced by formal consensus groups based upon the RAND design. We took as our starting point the common and direct approach of asking groups of clinicians to produce treatment recommendations based on their own experience and knowledge of the evidence. We investigated the effect on the extent of the group's disagreement of changing (1) provision of a review of the literature, (2) the specialty composition of the group, and (3) explicit recognition of the need to prioritise limited healthcare resources.

METHODS

We selected three conditions (chronic back pain, irritable bowel syndrome, and chronic fatigue syndrome) for study. They were chosen because: (1) although important problems, there were no national guidelines in the UK; (2) there was a mismatch between current clinical practice in the UK and the available research evidence; and (3) care was provided by at least two groups of clinicians (including general practitioners (GPs) and mental health professionals (MHPs)). We conducted a systematic review of the evidence on the effectiveness of mental health interventions in primary care for patients with these conditions.[17] Four relevant interventions were identified: behavioural therapy, cognitive behavioural therapy, brief psychodynamic interpersonal therapy, and antidepressants. A questionnaire comprising 64 clinical scenarios was then developed to elicit participants' judgments about the appropriate use of the four interventions for the three conditions studied[4 15] in the presence or absence of clinical and social situations (cues) identified as relevant by GPs and psychiatrists. The cues were: co-existent depressive

Abbreviations: GP, general practitioner; MADM, mean absolute deviation from the median; MHP, mental health professional

Qual Saf Health Care: first published as 10.1136/qshc.2004.013227 on 1 August 2005. Downloaded from http://qualitysafety.bmj.com/ on January 27, 2023 by guest. Protected by copyright.

symptoms, back pain induced insomnia (chronic back pain only), a financial motivation to return to work (chronic fatigue syndrome only), and patient belief that their condition had an organic cause. The appropriateness of each intervention with respect to physical and psychological outcomes was rated separately because the research evidence reported some differences in these respects. A total of 128 ratings were required (for example, ''When the patient suffering from irritable bowel syndrome demonstrates symptoms of depression, cognitive behavioural therapy is a good treatment option to improve physical functioning'') for an adult primary care patient of working age who exhibits enough features of the condition for the panellist to be satisfied with the diagnosis and who has no other symptoms. Ratings were made on Likert scales from 1 (strong disagreement with the scenario) to 9 (strong agreement).

Eight types of groups were established in a factorial design (fig 1) to allow every combination of three design factors: literature review (provided versus not provided); group composition (GP only versus a mixed group); and, as a first step towards bringing in the wider questions of cost and values, assumptions about the level of available healthcare resources (ideal versus current provision in the National Health Service). Each type was replicated once, resulting in 16 groups.

Participants in the groups provided with a literature review were given the review before making their first round ratings. It included a description of the methodological limitations of the available research, and the findings were arranged so that the outcomes for each intervention were described in terms of their benefits and harms. The groups not provided with a literature review held their meetings before the review's publication.[17]

In terms of resource context, participants in the ''ideal'' groups made their ratings assuming the availability of: competent appropriately trained clinical psychologists, liaison psychiatrists and psychotherapists; multidisciplinary functional complaints clinics (with clinical psychologists and physiotherapists) where patients with chronic fatigue, irritable bowel syndrome and chronic back pain can be referred; skilled therapists in brief psychodynamic interpersonal therapy; and an ''in house'' general practice counsellor. There was also the freedom to choose to whom they refer patients; minimal waiting times for good quality services; no financial barriers that limit choice of therapy; and timely and detailed feedback to the GP.

''Realistic'' or ''current'' resource provision assumed that patients can expect to wait for approximately 6 weeks to see the ''in house'' general practice counsellor; 3 months for an outpatient appointment to see a psychiatrist; 6 months for a psychology outpatient appointment (including clinical psychology, psychotherapy and counselling); and 6 weeks for an assessment at a pain clinic (followed by another wait of up to 6 months for treatment). There was also limited access to clinical psychologists, liaison psychiatrists and psychotherapists with expertise in managing patients with functional somatic symptoms; limited referral choices because services are organised on a geographical basis; variable quality of services in terms of their organisation and the range of clinical skills and experience that is available; and no chronic fatigue syndrome clinic.

The groups were established by selecting national random samples of GPs and MHPs from the Department of Health GP database for England (N = 27 723), the Royal College of Psychiatrists liaison section database, and the British Association of Behavioural and Cognitive Psychotherapists database (total N = 720).[4]

The aim was to establish 16 groups with 11 participants in each. The target number of participants in each group was based on evidence that, while large group sizes increased the reliability of group judgments, it could also make participation more unequal and increasing participant numbers above about 12 seldom increased the numbers of points of views expressed.[18] A total of 2680 GPs and 310 psychiatrists or other MHPs were initially invited to take part, assuming an initial response rate of 4% for GPs and 13% for MHPs. Invitations were sent 2 months before each meeting and the first 14 responders (stratified for the mixed groups) were recruited to allow for attrition.

Participants completed the first round of ratings by mail. Each group then met for a facilitated meeting at the same venue which followed a written protocol. The protocol included the instructions to be given to each group and a step by step guide to the process to be followed at the meeting. The first group was facilitated by NB and all the others by RR. The facilitator's role was to ensure that each scenario was discussed and that all participants had an equal chance to participate in the discussion. The facilitator reviewed the ratings before the meeting but did not lead the discussion or refer the group to any specific issues such as the research evidence. If asked for points of information, the facilitator redirected the question back to the group. At the meeting each participant was given a new copy of the questionnaire which included a reminder of their own initial ratings and the distribution of ratings for the group as a whole. Each scenario was discussed in turn and reasons for any differences explored, after which the participants privately re-rated the scenarios.

Analysis of data

The characteristics of participants (age, sex, and ethnicity) and their distribution across the groups were examined. Differences in participant characteristics between the mixed and GP only groups were tested using between-groups analysis of variance (ANOVA) for continuous variables and logistic regression clustered by group[19] for binary variables.

In formal consensus development methods the synthesis of individual opinions to produce recommendations on a particular question usually involves a measure of central location indicating the group's judgment as a whole, and a measure of the variation about this to indicate the extent of disagreement. In this study the group judgment for each scenario was given by the median of the group members' final ratings. The extent of disagreement for each scenario

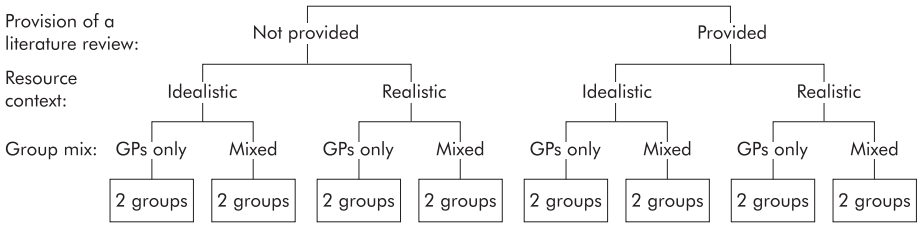

Figure 1 Study design.

Qual Saf Health Care: first published as 10.1136/qshc.2004.013227 on 1 August 2005. Downloaded from http://qualitysafety.bmj.com/ on January 27, 2023 by guest. Protected by copyright.

Table 1 Characteristics of groups and participants

	GP only groups (n = 8)	Mixed groups (n = 8)	Baseline differences*
Target number of participants	14 per group	14 per group	
Mean (range) no of participants per group	10.6 (9–14)	11.5 (10–13)	
Mean (range) no of mental health professionals per group	0	6.3 (5–7)	
Mean age in years (group range)	47.0 (42.2–51.9)	46.9 (43.2–49.8)	p = 0.82
Sex: % female (group range)	33.8 (11.1–54.5)	42.1 (9.1–63.6)	p = 0.01
Ethnicity: % non-white (group range)	17.1 (7.7–33.3)	14.0 (0.0–25.0)	p = 0.32

*p values from logistic regression clustered by group or ANOVA F tests between groups.

was given by the group's mean absolute deviation from the median (MADM)—that is, the average distance (on the 9-point Likert scale) of the participants' ratings from the group's median rating. The "classic" RAND definition of disagreement for a nine member panel was also calculated using the algorithm for panels of different sizes.[20]

We evaluated the impact of factors on the extent of disagreement using analysis of variance (ANOVA). The effects of three design factors were tested between groups and the effects of the clinical and social cues were tested within groups, with each group treated as a random effect. Interactions between the three design factors were tested at the $p<0.05$ level and interactions between design factors and the clinical and social cues were tested at a Bonferroni corrected level of $p<0.006$. The 95% confidence intervals were estimated for prespecified comparisons—for example, for scenarios involving patients with depression versus those with no signs of depression—from a random effects regression model for the group design factors and from a regression analysis clustered by group for the clinical and social cues.

Finally, disagreement might be greater towards the middle of the appropriateness scale because of "floor" and "ceiling" effects so we examined the relationship between the extent of disagreement in the groups and their median appropriateness ratings. We treated median ratings of 6.5–9 as representing the group view that an intervention was appropriate, ratings of 4–6 as equivocal, and ratings of 1–3.5 as inappropriate. We compared the three categories of appropriateness before and after adjusting for group design factors, treatment, and condition using a regression model with each group included as a random effect.

The study was approved by the London School of Hygiene and Tropical Medicine ethics committee.

RESULTS

Recruitment and response rates

A total of 135 GPs and 42 MHPs participated in the consensus groups. The groups ranged in size from 9 to 14 participants due to different group attrition rates (table 1). The mixed groups included a higher proportion of female participants

Table 2 Extent of disagreement by group design factors and clinical and social cues (lower values indicate closer consensus)

		Mean absolute deviation from the median (MADM)			
		Round 1 mean	Round 2 mean	Round 2 difference (95% CI)	ANOVA* F test (p value)
Group design factors					
Systematic literature review	Provided	1.32	1.15		
	Not provided	1.35	1.19	0.04 (−0.11 to 0.19)	$F_{1,12}=0.31$ (p = 0.59)
Group composition	GPs only	1.31	1.14		
	GPs and MHPs	1.36	1.21	0.07 (−0.08 to 0.22)	$F_{1,12}=0.89$ (p = 0.36)
Healthcare resource context	Assumption of "realistic" context	1.31	1.14		
	Assumption of "idealistic" context	1.37	1.20	0.06 (−0.09 to 0.20)	$F_{1,12}=0.24$ (p = 0.63)
Clinical and social cues					
Depression	No co-existent depressive symptoms	1.35	1.21		
	Co-existent depressive symptoms	1.27	1.06	−0.15 (−0.24 to −0.07)	
Organic cause	Patient willing to try any treatment	1.33	1.18		
	Patient believes condition has an organic cause	1.42	1.27	0.09 (0.01 to 0.18)	
Financial motivation to return to work	No	1.31	1.20		$F_{7,105}=7.26$ (p<0.001)
	Yes	1.36	1.12	−0.08 (−0.19 to 0.03)	
Back pain induced insomnia	No	1.30	1.11		
	Yes	1.32	1.13	0.02 (−0.07 to 0.10)	
Condition	Chronic back pain	1.31	1.13		
	Irritable bowel syndrome	1.38	1.24	0.11 (0.01 to 0.21)	$F_{2,30}=3.90$ (p = 0.03)
	Chronic fatigue syndrome	1.34	1.16	0.03 (−0.05 to 0.12)	
Treatment	Behavioural therapy	1.20	1.06		
	Cognitive behavioural therapy	1.27	1.07	0.01 (−0.07 to 0.08)	$F_{3,45}=18.89$ (p<0.001)
	Brief psychodynamic interpersonal therapy	1.37	1.22	0.16 (0.09 to 0.23)	
	Antidepressants	1.51	1.34	0.28 (0.21 to 0.34)	

*From the single ANOVA model without interaction terms (all design factor interactions p>0.05, all design factor:clinical and social cue interactions p>0.006).

Qual Saf Health Care: first published as 10.1136/qshc.2004.013227 on 1 August 2005. Downloaded from http://qualitysafety.bmj.com/ on January 27, 2023 by guest. Protected by copyright.

Table 3 Difference in extent of disagreement by group rating of appropriateness

Group rating of appropriateness	Mean consensus*	Difference (95% CI)	Adjusted† difference (95% CI)
Appropriate	1.03	−0.27 (−0.24 to −0.29)	−0.23 (−0.20 to −0.26)
Equivocal	1.29	–	–
Not appropriate	1.33	0.03 (−0.01 to 0.07)	−0.01 (−0.05 to 0.03)

*Measured as the mean absolute deviation from the median.
†Adjusted for group design factors, treatment and condition.

than the GP only groups because a higher proportion of MHPs were female (50.0% *v* 34.8% of GPs).

Extent of disagreement

The mean (range) of the MADMs was 1.34 (0.33–2.91) in round 1 and 1.17 (0.20–2.55) in round 2. For all factors there was a reduction in the mean MADM between rounds (table 2).

Methodological factors

There was no evidence that group composition, provision of a literature review, or assumptions about the level of healthcare resources were associated with differences in the extent of disagreement (table 2). There was also no evidence for interactions between these design factors and the clinical and social cues (all $p>0.006$).

Clinical and social factors

There was strong evidence for variations in the extent of disagreement between treatments ($p<0.001$) and social cues ($p<0.001$). The consensus was closer for behavioural therapy and cognitive behavioural therapy than for brief psychodynamic interpersonal therapy and antidepressants. There was also closer consensus about treatment for patients with symptoms of depression than for patients without symptoms of depression, and for patients who would try anything (that is, accepted that their condition might have a psychological basis) than for patients who believed it was organic. There was some evidence for differences in the closeness of consensus between conditions ($p = 0.03$), although the only significant pairwise comparison ($p<0.05$) was the closer consensus for back pain than irritable bowel syndrome.

Group view of appropriateness of intervention

If groups rated an intervention appropriate, the consensus tended to be closer than if they rated it equivocal or inappropriate (table 3).

Comparison with the ''classic'' RAND definition of disagreement

In round 1 there were 149 (7.3%) scenarios rated ''with disagreement'', 17 (0.8%) of which would involve altering a judgment from appropriate or inappropriate to uncertain. The corresponding figures for round 2 were 84 (4.1%) and 6 (0.3%). The MADMs for the scenarios rated ''with disagreement'' ranged from 1.40 to 2.91.

DISCUSSION

The provision of a literature review and differences in group composition or assumptions about resources did not significantly affect the extent of disagreement. Consensus was greater where groups agreed that a treatment was appropriate. The extent of consensus was also related to the treatment offered and to patient characteristics (the presence of depression and readiness to try any treatment).

Methodological issues

A major strength of this study is its design. To evaluate whether literature review, group composition, or healthcare resource context affects the extent of consensus, the appropriate unit of analysis is the group.[21] By including 16 groups, with each design replicated, we were able to analyse our data in this way and, by using a random effects analysis, to treat our groups as a sample drawn from all possible groups. This allowed us to take into account correlation between the multiple scenarios rated by a group and the fact that groups of the same design do not produce identical results.[22 23]

In common with previous studies,[24–28] we used the MADM rather than the standard deviation, for two reasons. Firstly, it does not give extra weight to extreme observations and, secondly, it measures variation about the median which is the most commonly used measure of central tendency in consensus development. We did not use the interquartile range because it lacks the sensitivity of the MADM, even though its lack of sensitivity to extreme values could be an advantage in some cases. There are several other ways of combining the ratings of individual participants into a single measure of degree of consensus within the group. The RAND approach is perhaps the best known and involves defining ''agreement'' or ''disagreement'' based on the number of participants with ratings outside a specified range. However, this dichotomous approach gives different results depending on the size of the group[20 29] and how ''outliers'' or extreme ratings are treated.[27 28 30] Comparison with the ''classic'' RAND definition of disagreement indicated that, in the second round, 4.1% of our scenarios would be rated ''with disagreement''. We used regression clustered by group to calculate 95% confidence intervals for the clinical and social cues because it was slightly more conservative than random effects regression and was consistent with the ANOVA. Both approaches produced identical effect sizes.

Our study differed from many guideline panels in that our panel members were randomly selected because we wanted unbiased estimates of our design effects. Also, our panel members were drawn from whole populations of practitioners who work with patients with these conditions as part of their daily practice rather than selected on the grounds of specific expertise, experience, or reputation. Given the large numbers of panels in our design, this was the only practicable approach. It is possible that recognised experts would have been less inclined to change their views and more inclined to overstate the effectiveness of their own interventions.[6 7] In practice, contributions to guideline development are generally sought from practising clinicians as well as from experts, with practitioners receiving incentives to participate in the form of credits for continuous professional development.

Finally, these findings may not be generalisable to conditions with fewer psychosocial determinants and a clearer pathogenesis.

Relationship with previous studies

There has been no previous research on the impact of providing a systematic review on the closeness of consensus.

Qual Saf Health Care: first published as 10.1136/qshc.2004.013227 on 1 August 2005. Downloaded from http://qualitysafety.bmj.com/ on January 27, 2023 by guest. Protected by copyright.

We have previously reported how the groups that were provided with a literature review produced judgments that were nearer to the research evidence,[4] and one might have expected to see closer consensus around these judgments too, but this did not occur. There was a shift in the distribution of participants' ratings but no change in the spread of that distribution.

Previous research has shown how individuals from different specialties differ in their ratings. This might be expected to result in less consensus from mixed groups than from single specialty groups, but we did not find this despite the fact that the mixed groups produced judgments rating brief psychodynamic interpersonal therapy and antidepressants less favourably than the GP only groups. This may be because GPs are not homogenous in their ratings or because, in our mixed groups, GPs and MHPs took account of each others' opinions.[4]

Previous research[5] has suggested that differences in judgments and in the extent of consensus between panels from different countries may be due to differences in healthcare resources. However, we found that assumptions about the healthcare resources available had no significant effect on the closeness of consensus. This may be because the participants were more concerned—and perhaps more familiar—with judging effectiveness than cost or cost effectiveness. Also, they may have found it difficult to hypothesise an ''ideal'' situation.[4]

There has been little consideration of the relationship between the point on a Likert scale of a group judgment and the closeness of group consensus around that judgment. One plausible hypothesis is that this relationship is symmetrical, with closer consensus at the extremes of the scale[20] (strong agreement or strong disagreement). We found that, although closer consensus occurred at one end of the scale (where groups rated the interventions appropriate), there was no corresponding effect at the other end (rating an intervention inappropriate). Our findings are consistent with an asymmetry in the attitudes of some of our participants, who may have been more willing to give unequivocal support than unequivocal condemnation.

On the particular topics considered, closer consensus may have been achieved for chronic back pain because of the widespread provision of multidisciplinary clinics which provide psychological interventions for such patients.[31] Clinicians were therefore likely to be more familiar with the concept of using psychological interventions such as cognitive behavioural therapy and behavioural therapy for chronic back pain than for irritable bowel syndrome where pharmacotherapy is the norm.

Implications

In an earlier analysis we found that group judgments were affected by provision of a literature review and group composition, but not by assumptions about resources.[4] In this analysis we found that the extent of disagreement around those judgments was not affected by any of these factors. In some ways these results are encouraging because the extent of agreement or consensus among participants in consensus processes appears to be sensitive to scenarios and research results but robust to variations in the design of the process. Our findings illustrate the importance of presenting the group judgments and the extent of agreement around those judgments separately when clinical guidelines are published. For example, the knowledge that a recommendation is underpinned by a high degree of agreement in a guideline development group may engender higher levels of adherence by practising clinicians than one for which there is less agreement.

ACKNOWLEDGEMENTS

This study is part of a research programme examining the methodology of group decision making for clinical guideline development. The authors thank Tom Sensky, Andy Haines, Simon Carter and Theresa Marteau for advice; the GPs and mental health specialists who participated in the study; and Kirsten Larkin for assistance.

Authors' affiliations

A Hutchings, R Raine, C Sanderson, N Black, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

The research programme is funded by a Medical Research Council (MRC) Clinician Scientist Fellowship for Rosalind Raine. The MRC was the sole funding source and they had no involvement in study design, collection, analysis and interpretation of data, in writing or submitting the paper.

All authors declare that they have no competing interests.

REFERENCES

1 **Woolf SH**, Grol R, Hutchinson A, *et al*. Potential benefits, limitations and harms of clinical guidelines. *BMJ* 1999;**318**:527–30.
2 **Burgers JS**, Grol R, Klazinga NS, *et al*. Towards evidence-based clinical practice: an international survey of 18 clinical guideline programmes. *Int J Qual Health Care* 2003;**15**:31–45.
3 **Savoie I**, Kazanjian A, Bassett K. Do clinical practice guidelines reflect research evidence? *J Health Serv Res Policy* 2000;**5**:76–82.
4 **Raine R**, Sanderson C, Hutchings A, *et al*. An experimental study of determinants of group judgments in clinical guideline development. *Lancet* 2004;**354**:429–37.
5 **Murphy MK**, Black NA, Lamping DL, *et al*. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;**2**(3):1–88.
6 **Herrin J**, Etchason JA, Kahan JP, *et al*. Effect of panel composition on physician ratings of appropriateness of abdominal aortic aneurysm surgery: elucidating differences between multispecialty panel results and specialty society recommendations. *Health Policy* 1997;**42**:67–81.
7 **Fitch K**, Lázaro P, Aguilar MD, *et al*. Physician recommendations for coronary revascularization: variations by clinical specialty. *Eur J Public Health* 1999;**9**:181–7.
8 **Landrum MB**, McNeil BJ, Silva L, *et al*. Understanding variability in physician ratings of the appropriateness of coronary angiography after acute myocardial infarction. *J Clin Epidemiol* 1999;**52**:309–19.
9 **Bernstein SJ**, Lazaro P, Fitch K, *et al*. Effect of specialty and nationality on panel judgments of the appropriateness of coronary revascularization: a pilot study. *Med Care* 2001;**39**:513–20.
10 **Brook RH**. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel RA, eds. *Clinical practice guidelines development: methodology perspectives*. Rockville, MD: Public Health Service, US Department of Health and Human Services, 1994:59–70.
11 **Campbell SM**, Hann M, Roland MO, *et al*. The effect of panel membership and feedback on ratings in a two-round Delphi survey: results of a randomized controlled trial. *Med Care* 1999;**37**:964–8.
12 **Brook RH**, Kosecoff JB, Park RE, *et al*. Diagnosis and treatment of coronary disease: comparison of doctors' attitudes in the USA and the UK. *Lancet* 1988;i:750–3.
13 **Vader JP**, Porchet F, Larequi-Lauber T, *et al*. Appropriateness of surgery for sciatica: reliability of guidelines from expert panels. *Spine* 2000;**25**:1831–6.
14 **Vader JP**, Burnand B, Froehlich F, *et al*. Appropriateness of upper gastrointestinal endoscopy: comparison of American and Swiss criteria. *Int J Qual Health Care* 1997;**9**:87–92.
15 **Burnand B**, Vader JP, Froehlich F, *et al*. Reliability of panel-based guidelines for colonoscopy: an international comparison. *Gastrointest Endosc* 1998;**47**:162–6.
16 **Cluzeau FA**, Littlejohns P, Grimshaw JM, *et al*. Development and application of a generic methodology to assess the quality of clinical guidelines. *Int J Qual Health Care* 1999;**11**:21–8.
17 **Raine R**, Haines A, Sensky T, *et al*. Systematic review of mental health interventions for patients with common somatic symptoms: can research evidence from secondary care be extrapolated to primary care? *BMJ* 2002;**325**:1082–5.
18 **Richardson FM**. Peer review of medical care. *Med Care* 1972;**10**:29–39.
19 **StataCorp**. *Stata statistical software: release 8.1*. College Station, TX: Stata Corporation, 2003.
20 **Fitch K**, Bernstein SJ, Aguilar MD, *et al*. *The RAND/UCLA appropriateness method user's manual*. Santa Monica, CA: RAND, 2001.
21 **Wood J**, Freemantle N. Choosing an appropriate unit of analysis in trials of interventions that attempt to influence practice. *J Health Serv Res Policy* 1999;**4**:44–8.
22 **Shekelle PG**, Kahan JP, Bernstein SJ, *et al*. The reproducibility of a method to identify the overuse and underuse of medical procedures. *N Engl J Med* 1998;**338**:1888–95.

Qual Saf Health Care: first published as 10.1136/qshc.2004.013227 on 1 August 2005. Downloaded from http://qualitysafety.bmj.com/ on January 27, 2023 by guest. Protected by copyright.

23 **Coulter ID**, Marcus M, Freed JR. Consistency across panels of ratings of appropriateness of dental care treatment procedures. *Community Dent Health* 1998;**15**:97–104.
24 **Park RE**, Fink A, Brook RH, *et al. Physician ratings of appropriate indications for six medical and surgical procedures*. Santa Monica, CA: RAND, 1986.
25 **Shekelle P**, Schriger DL. Evaluating the use of the appropriateness method in the Agency for Health Care Policy and Research clinical practice guideline development process. *Health Serv Res* 1996;**31**:453–68.
26 **Vella K**, Goldfrad C, Rowan K, *et al.* Use of consensus development to establish national research priorities in critical care. *BMJ* 2000;**320**:976–80.
27 **Scott EA**, Black N. When does consensus exist in expert panels? *J Public Health Med* 1991;**13**:35–9.
28 **Imamura K**, Gair R, McKee M, *et al.* Appropriateness of total hip replacement in the United Kingdom. *World Hosp Health Serv* 1997;**32**:10–4.
29 **Naylor CD**, Basinski A, Baigrie RS, *et al.* Placing patients in the queue for coronary revascularization: evidence for practice variations from an expert panel process. *Am J Public Health* 1990;**80**:1246–52.
30 **Park RE**, Fink A, Brook RH, *et al.* Physician ratings of appropriate indications for three procedures: theoretical indications vs indications used in practice. *Am J Public Health* 1989;**79**:445–7.
31 **Guzman J**, Esmail R, Karjalainen K, *et al.* Multidisciplinary bio-psycho-social rehabilitation for chronic low back pain. *Cochrane Database Syst Rev* 2002;**1**:CD000963.

bmjupdates+

bmjupdates+ is a unique and free alerting service, designed to keep you up to date with the medical literature that is truly important to your practice.
bmjupdates+ will alert you to important new research and will provide you with the best new evidence concerning important advances in health care, tailored to your medical interests and time demands.

Where does the information come from?
bmjupdates+ applies an expert critical appraisal filter to over 100 top medical journals
A panel of over 2000 physicians find the few 'must read' studies for each area of clinical interest

Sign up to receive your tailored email alerts, searching access and more…

www.bmjupdates.com

Qual Saf Health Care: first published as 10.1136/qshc.2004.013227 on 1 August 2005. Downloaded from http://qualitysafety.bmj.com/ on January 27, 2023 by guest. Protected by copyright.