Journal of Health Services Research & Policy
http://hsr.sagepub.com/

A comparison of formal consensus methods used for developing clinical guidelines

Andrew Hutchings, Rosalind Raine, Colin Sanderson and Nick Black

J Health Serv Res Policy 2006 11: 218
DOI: 10.1258/135581906778476553

The online version of this article can be found at:
http://hsr.sagepub.com/content/11/4/218

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for *Journal of Health Services Research & Policy* can be found at:

Email Alerts: http://hsr.sagepub.com/cgi/alerts

Subscriptions: http://hsr.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2006

What is This?

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

A comparison of formal consensus methods used for developing clinical guidelines

Andrew Hutchings, Rosalind Raine, Colin Sanderson, Nick Black
Health Services Research Unit, London School of Hygiene & Tropical Medicine, London, UK

***Objectives*: To compare two consensus development methods commonly used for developing clinical guidelines in terms of the judgments produced, closeness of consensus, amount of change between rounds, concordance with research evidence and reliability.**

***Methods*: In all, 213 general practitioners and mental health professionals from England participated in four Delphi and four nominal groups. They rated the appropriateness of four treatments (cognitive behavioural therapy [CBT], behavioural therapy [BT], brief psychodynamic interpersonal therapy [BPIT] and antidepressants) for three conditions. First, participants rated the appropriateness of interventions independently, using a postal questionnaire. For nominal groups, the ratings were fed back and discussed at a meeting, and then group members privately completed the questionnaire again. For Delphi groups, there was feedback but no discussion, and the entire process was conducted by postal questionnaire.**

***Results*: The effect of consensus method on final ratings varied with therapeutic intervention, with nominal groups rating CBT and antidepressants more favourably than Delphi groups. Consensus was closer in the nominal than in the Delphi groups in both rounds. There was no overall difference between groups in their concordance with research evidence (odds ratio 1.13, 95% confidence interval 0.79–1.61). In this study, the Delphi method was more reliable (κ coefficients 0.88 and 0.89 compared with 0.41 and 0.65 for nominal groups).**

***Conclusions*: The advantages of nominal groups (more consensus; greater understanding of reasons for disagreement) could be combined with the greater reliability of the Delphi approach by developing a hybrid method.**

Journal of Health Services Research & Policy Vol 11 No 4, 2006: 218–224 © The Royal Society of Medicine Press Ltd 2006

Introduction

Clinical practice guidelines are extensively used to improve the quality of health care. The methods used for developing guidelines vary but are increasingly based on a combination of the best available scientific evidence and consensus judgments obtained by formal, explicit methods.[1,2]

Consensus methods typically require experts to make individual judgments in private before, and then again after, exposure to the views of other group members. In the modified nominal group (NG) technique, participants first express their views independently using a postal questionnaire. They then meet for a facilitated discussion after which they complete the questionnaire again privately.[3] The main alternative is the postal Delphi survey, in which the summarized results of successive rounds are sent back to the participants so that they can revise their opinions if they wish.[4] This allows larger groups, but the opportunities for clarification and resolution of disagreements are more limited than in the modified NG since participants never meet.

There have been few comparisons of modified nominal and Delphi methods for the development of clinical guidelines. Four studies found that the methods produced similar final ratings[5–7] and similar levels of within-group agreement.[8] No studies have investigated differences in the change in ratings between rounds – a crucial research gap in view of the differing processes. The extent to which NG ratings agree with research evidence has been reported to be moderate,[9,10] but no comparison of Delphi ratings with the evidence has been reported. The reliability of NGs of similar composition was moderate or good.[11–13] The reliability of the Delphi method was also good.[4] However, the reliability of the NG compared with the Delphi method for clinical guideline development has not been examined.

Thus there is insufficient research evidence about the relative advantages of the two methods. Yet the choice of methods may affect the guidance published by national decision making bodies such as the National Institute for Health and Clinical Excellence (NICE) in England and Wales. Our aim was to compare nominal

Andrew Hutchings MSc, Lecturer in Health Services Research, **Rosalind Raine PhD**, MRC Clinician Scientist, **Colin Sanderson PhD**, Reader in Health Services Research, **Nick Black MD**, Professor of Health Services Research, Health Services Research Unit, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Correspondence to: andrew.hutchings@lshtm.ac.uk

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

and Delphi methods in terms of the overall group judgment, the extent of agreement within the groups, the change of judgment between rounds, concordance with the scientific evidence and the reliability of the methods (agreement between groups).

The data came from a larger research programme conducted in England that involved 16 NGs and four Delphi groups (Figure 1). In this programme, the way that NGs were conducted was varied (by altering group composition, whether or not a literature review was provided and assumptions about the resources available for health care). The impact of these variations has been reported elsewhere.[14,15]

Methods

Three conditions (chronic back pain, irritable bowel syndrome and chronic fatigue syndrome [CFS]) were selected.[15] We conducted a systematic review of the effectiveness of mental health interventions in primary care for patients with these conditions.[16] Four relevant interventions were identified: behavioural therapy (BT), cognitive behavioural therapy (CBT), brief psychodynamic interpersonal therapy (BPIT) and antidepressants.

A questionnaire covering 128 clinical scenarios was developed to elicit the views of participating general practitioners (GPs) and mental health professionals (MHPs) about the appropriateness of the four interventions for the three conditions, in the presence or absence of four clinical and social situations (cues), previously identified by GPs and psychiatrists. These cues were (i) coexistent depressive symptoms; (ii) clinicians' perception that the patient believes that their condition has an organic cause; (iii) insomnia in patients with chronic back pain and (iv) a financial motivation to return to work in patients with CFS. For example, one scenario was the use of behavioural therapy for improving physical outcomes in a patient with CFS who believed their condition had an organic cause. Participants rated their level of agreement for each scenario on Likert scales where 1 = strong disagreement and 9 = strong agreement.[15]

The GPs and MHPs were randomly selected from professional databases and invited to participate (Figure 2). For each of the 16 NGs, we recruited 14 participants, aiming for 11 in each[17] after attrition. Eight of the NGs were clinically homogeneous (GPs only) and eight were a mixture of GPs and MHPs. GPs and MHPs who agreed to participate in the Delphi groups (two with GPs only and two mixed) and returned with completed first round ratings were randomly allocated to one of four groups (46 per group).

Both types of group completed the first round of ratings by post. For the second round, each NG met for a facilitated meeting, which followed a written protocol.[14] At the meeting, each participant was given feedback in the form of the distribution of ratings for the whole group with a remainder of their own initial ratings. Each scenario was discussed in turn and reasons for any differences were explored. The participants then privately re-rated each scenario. Delphi group participants were sent feedback in the same format as the NGs, and returned their second round ratings by post.

This paper is based on data from all four Delphi groups and their matching NGs (Figure 1). All participants were sent a literature review, and all ratings were made in the context of realistic levels of health care resources for England.[15] The characteristics of participants were compared using χ^2- and t-tests.

Ratings at each round

A group's rating for a scenario was defined as the median of the participants' ratings on the nine-point Likert scale. Mean differences in median ratings between Delphi and NGs were calculated using linear regression estimated by maximum likelihood. Groups were included as random effects to allow for the large number of scenarios rated (Appendix). Semi-robust

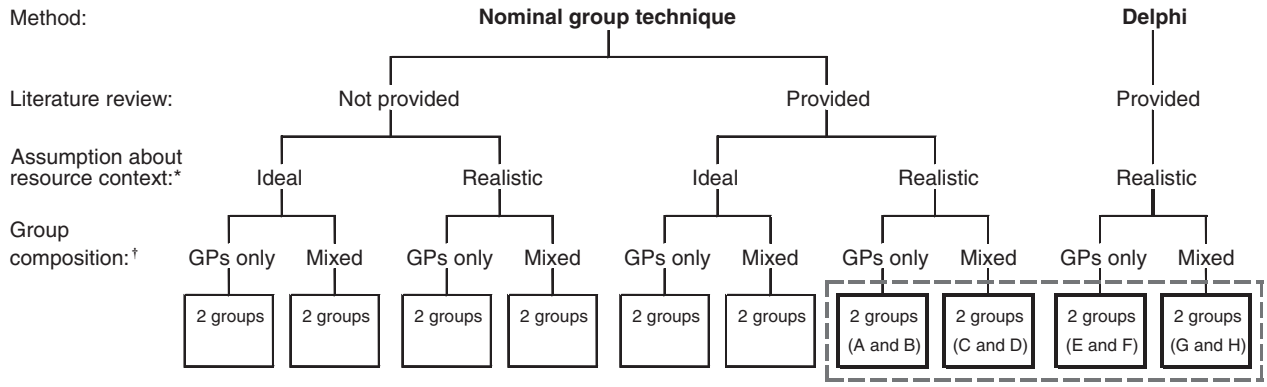

Figure 1 Study design.
Dotted lines round boxes indicate groups included in the current analyses. *Ideal resource context was defined as the immediate availability of appropriately trained, relevant health professionals. Realistic resource context referred to the existence of defined waiting lists and limited referral choice.[21] †The mixed groups included general practitioners and mental health professionals (psychiatrists, psychologists, counsellors and mental health nurses)

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

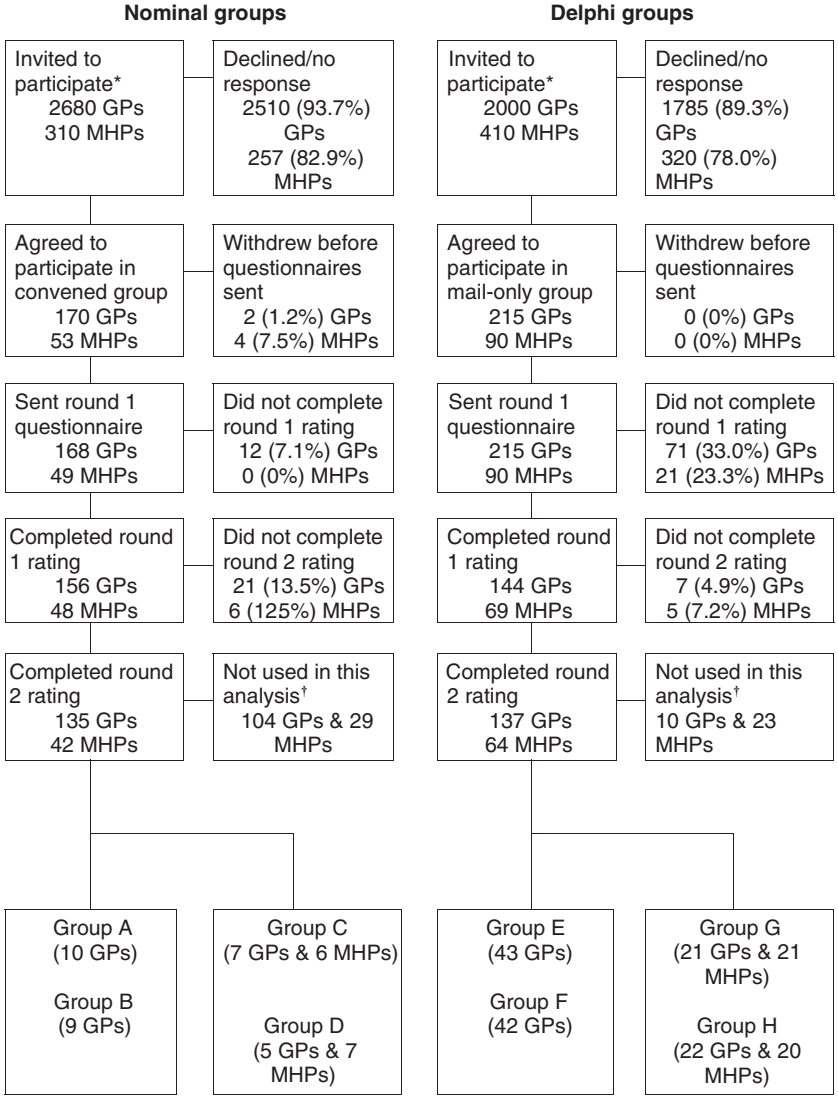

Figure 2 Recruitment of study participants.
*Randomly selected from the Department of Health GP database for England ($N = 27{,}723$), the Royal College of Psychiatrists liaison section database and the British Association of Behavioural and Cognitive Psychotherapists database (total $N = 720$). †12 additional nominal groups of different designs were created but not included in this comparison because there were no matching mail-only groups. ‡Because of the response rate, these participants were allocated to additional groups

standard errors were used to relax assumptions about constant variance between groups of each design. Treatments were considered separately because we have previously found evidence for effect modification between treatments and group-mix.[15]

Changes in ratings between rounds

Differences between Delphi and NGs in the *changes* in ratings between rounds were assessed by repeating the group-level analysis for the round 2 ratings, with round 1 ratings included as a covariate.

Extent of within-group agreement

The mean absolute deviation from the median (MADM) of the participants' Likert scale ratings was used as an indicator of the extent of agreement for a scenario within a group. Mean differences in the MADMs were estimated using the same approach as for the ratings.

Concordance with research evidence

Logistic regression was used to assess the effect of the method on the number of group ratings agreeing with the evidence (good evidence of benefit in primary care, good evidence of no benefit in primary care, no clear evidence).[16]

Inter-group reliability

The reliability of each method was assessed by measuring agreement of median ratings between pairs of groups of the same design using κ-coefficients.

The study received ethical approval from the London School of Hygiene & Tropical Medicine Ethics Committee.

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

Results

Participation

Participation rates are given in Figure 2. For the eight groups examined in this study, there were no significant differences between the 44 NG and 169 Delphi participants by sex (men 61.4 versus 66.3%, $P = 0.54$), ethnicity (non-white 18.2 versus 13.7%, $P = 0.45$), mean age (45.8 versus 45.2 years, $P = 0.66$), profession (GP 70.5 versus 75.7%, $P = 0.47$) or residence in London and the southeast versus the rest of England (43.2 versus 42.6%, $P = 0.95$).

Comparison of final ratings and extent of agreement

Differences in final ratings between nominal and Delphi groups depended on the treatment being considered (Table 1). For BT and CBT, there was no significant difference, whereas antidepressants were rated as more appropriate by the nominal than by the Delphi groups. The largest mean difference was for BPIT, but the wide confidence interval reflected inconsistent ratings between the nominal groups (Figure 3). Within-group agreement was closer in the

Table 1 Group judgments and extent of within-group agreement for each therapeutic intervention and for rounds 1 and 2

	Round 1			Round 2			Change between rounds*
	Nominal groups	Delphi groups	Difference (95% CI)	Nominal groups	Delphi groups	Difference (95% CI)	Difference (95% CI)
Group judgments†							
CBT	6.78	6.54	0.24 (−0.28 to 0.76)	6.98	6.67	0.31 (−0.14 to 0.76)	0.12 (−0.02 to 0.26)
BT	6.14	6.17	−0.02 (−0.42 to 0.37)	6.23	6.28	−0.05 (−0.44 to 0.35)	−0.03 (−0.26 to 0.21)
BPIT	4.46	3.71	0.75 (0.12 to 1.37)	4.73	3.71	1.02 (−0.19 to 2.22)	0.42 (−0.33 to 1.18)
Antidepressants	5.16	4.82	0.34 (−0.05 to 0.72)	5.43	4.81	0.61 (0.26 to 0.96)	0.31 (0.02 to 0.60)
Extent of within-group agreement‡							
CBT	1.15	1.34	−0.19 (−0.42 to 0.04)	0.93	1.11	−0.18 (−0.35 to −0.02)	−0.05 (−0.08 to −0.01)
BT	1.19	1.37	−0.17 (−0.34 to −0.01)	1.00	1.14	−0.14 (−0.31 to 0.03)	−0.02 (−0.12 to 0.08)
BPIT	1.41	1.63	−0.22 (−0.39 to −0.04)	1.17	1.51	−0.33 (−0.46 to −0.20)	−0.21 (−0.34 to −0.08)
Antidepressants	1.40	1.57	−0.17 (−0.34 to −0.01)	1.26	1.38	−0.12 (−0.36 to 0.12)	0.01 (−0.13 to 0.14)

*Nominal groups compared with Delphi groups from analysis of covariance
†Mean of group ratings of appropriateness
‡Mean absolute deviation from the median (smaller values indicate greater within-group agreement)

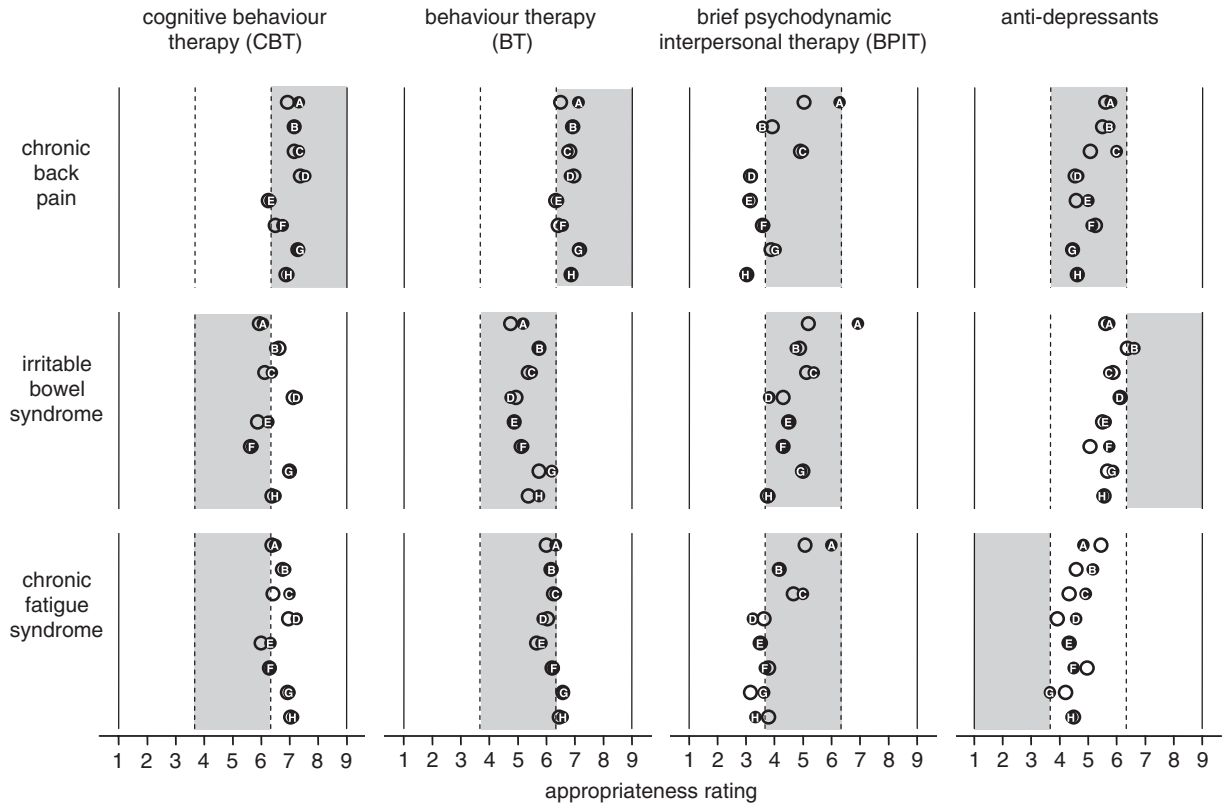

Figure 3 Change in group judgments by intervention and condition.
○ Round 1 mean of each group's ratings of all scenarios for the intervention and condition. ● Round 2 mean of each group's ratings of all scenarios for the intervention and condition. The shaded areas represent our assessment of the evidence for effectiveness in primary care.[23] (A, B) The two GP-only nominal groups. (C, D) The two mixed nominal groups. (E, F) The two GP-only Delphi groups. (G, H) The two mixed Delphi groups

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

nominal than in the Delphi groups in both rounds, although significantly so only for BT, BPIT and antidepressants in round 1, and CBT and BPIT in round 2.

Effect of consensus method on change in ratings and extent of agreement

For BT and CBT, there was no significant difference between the nominal and Delphi groups in how much ratings changed (Table 1). For antidepressants, NGs had a larger shift towards more favourable ratings. For BPIT, the NGs changed their ratings in different directions (Figure 3). At the individual level also, NG participants were more likely to change their ratings of scenarios between rounds than Delphi participants (38.7 versus 28.4% of scenarios changed, $P<0.001$). There were significant differences in the change in within-group agreement for CBT and BPIT, for which agreement increased more in the NGs.

Concordance with research evidence

At round one, overall concordance with our assessment of the research evidence was greater in NGs. At round two, the two types of group were similar in this respect (Table 2). However, as a result of NGs' tendency to rate CBT and antidepressants more favourably than Delphi groups, the NGs were more concordant than Delphi after round 2 for scenarios with evidence of benefit (e.g. CBT in chronic back pain) and less concordant for scenarios with evidence of no benefit (i.e. antidepressants in CFS).

Reliability

Between-group agreement for round one ratings was closer for the pairs of Delphi groups than for the NGs (Table 3). Agreement between the NGs was poorer in round 2 than round 1, particularly for the GP-only groups. There was no change in agreement between the Delphi groups.

Discussion

This is the first study to compare nominal with Delphi groups using more than one group of each type. The NGs rated antidepressants more favourably than the Delphi groups, which explained why overall NGs agreed more closely with evidence of effectiveness

Table 2 Concordance of group ratings with the research evidence

		Nominal groups (% concordance)		Delphi groups (% concordance)		Odds ratio for concordance (95% confidence interval)†	
Research evidence*	**Consensus**	**Round 1**	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Benefit‡ (32 scenarios)	Benefit	75.0	75.8	57.3	64.1	2.27 (1.09 to 4.75)	1.74 (0.72 to 4.21)
	Unclear	24.2	24.2	41.4	35.9		
	No benefit	0.8	0.0	1.6	0.0		
Unclear§ (84 scenarios)	Benefit	23.2	33.0	19.6	23.8	1.60 (0.83 to 3.06)	1.16 (0.65 to 2.06)
	Unclear	68.5	56.3	58.3	52.7		
	No benefit	8.3	10.7	22.0	23.5		
No benefit** (12 scenarios)	Benefit	16.7	16.7	12.5	12.5	1.11 (0.46 to 2.70)	0.27 (0.08 to 0.84)
	Unclear	52.1	66.7	58.3	45.8		
	No benefit	31.2	16.7	29.2	41.7		
Concordant overall		66.6	57.5	55.3	54.5	1.62 (1.25 to 2.08)	1.13 (0.79 to 1.61)

*Evidence of effectiveness in primary care. Two researchers (RR and KL) independently categorized the evidence as: good evidence of benefit in primary care (a statistically significant improvement at least six months post-intervention); good evidence of no benefit in primary care (no statistically significant improvement in studies totalling at least 100 patients) and no clear evidence (all remaining situations)[16]. Group median ratings of 6.5 to 9.0, 1.0 to 3.5, and 4.0 to 6.0 were taken as consistent with each of these categories, respectively
†In nominal compared with Delphi groups, adjusted for group composition (GP-only or mixed) and with groups included as random effects
‡All CBT and BT scenarios for back pain, all antidepressant scenarios for irritable bowel syndrome
§All BPIT and antidepressant scenarios for back pain, all CBT, BT and BPIT scenarios for irritable bowel syndrome, all CBT, BT and BPIT scenarios for chronic fatigue syndrome
**All antidepressant scenarios for chronic fatigue syndrome

Table 3 Between-group reliability of ratings using weighted kappa (K_w) for agreement

		K_w (95% CI)*	
Group design	**Group sizes**	**Round 1**	**Round 2**
Nominal (GP-only)	10, 9	0.67 (0.50 to 0.84)	0.41 (0.25 to 0.57)
Nominal (GPs & MHPs)	13, 12	0.77 (0.61 to 0.93)	0.65 (0.50 to 0.79)
Delphi (GP-only)	43, 42	0.88 (0.71 to 1.00)	0.88 (0.71 to 1.00)
Delphi (GPs & MHPs)	42, 42	0.90 (0.73 to 1.00)	0.89 (0.72 to 1.00)

*Quadratic-weighted κ-coefficient with confidence intervals calculated from the K_w standard error

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

and the Delphis more closely with evidence of ineffectiveness. NGs also demonstrated closer within-group agreement than Delphi groups, and the views of NGs shifted more between rounds than Delphi groups.

Delphi groups were more reliable than NGs. This was partly because of their larger size. However, the NGs' improvement in within-group agreement about BPIT in round 2 was accompanied by deterioration in between-group agreement (reliability). Given the absence of clear research evidence for BPIT, this may have been the result of exposure of different groups to different combinations of argument, anecdote and persuasive or dominating personalities.

Methodological considerations

Contributions to guideline development are increasingly being sought from clinicians with day-to-day experience of relevant patients. Consumers and policymakers may also be involved. For our study, participants were randomly selected from practitioners who work with these conditions as part of their daily practice. Given the large number of groups in our design and the three disparate conditions studied, this was the only practical approach, but may reduce the generalizability of our findings.

Our participation rates were low, as expected from the experience of the Medical Research Council GP Research Framework. Like every study, participants may differ from non-participants. The relevant issue here was whether the participants in the different groups (nominal versus Delphi) differed, and we found that they did not.

Commonly, analyses of consensus data reduce the nine-point scale to three categories: appropriate, inappropriate or uncertain.[18] Re-analysis of our data using these categories did not alter our conclusions.

Comparison with other studies

Previous studies report that the methods produce similar final ratings.[5–7] We found that differences depended on the treatment considered, and where differences did exist, NGs tended to rate more favourably than Delphi groups.

Our finding of greater within-group agreement in NGs is contrary to earlier research.[8] The difference was seen in the first round in our study and persisted. This may be due to our choice of topics. Arguably these were more controversial than previous studies,[19] possibly exacerbating any concerns in NGs about having to defend unorthodox opinions at a meeting.

Our finding that concordance with the research evidence varied according to the indication being rated is consistent with previous research.[9,10,15] The literature provides no other direct comparisons of the reliability of nominal and Delphi groups for developing clinical guidelines.

Implications

We chose clinical topics that were controversial in terms of both pathogenesis and choice of intervention. Caution is needed when drawing more general conclusions, but arguably it is in precisely these situations that guidelines on the basis of consensus are most useful.

Our results suggest that opinions may be more favourable to treatment, and are more likely to shift, when groups meet. This may or may not be an advantage. In theory, exploration of a topic should lead to a better understanding of the issues and 'better' results. On the other hand, we found that direct exposure to argument, anecdote and dominating personalities could lead different groups in different directions. Delphi groups are more reliable, partly because the group interaction is indirect and partly because more people can be involved.

The way forward for guideline production may be to draw upon the advantages of both methods by using a 'hybrid' approach. This could include a convened group in which the discussion is recorded to enable a thematic analysis of the issues aired. This would be followed by a postal stage in which the questionnaire and results, including the thematic analysis, from the convened group would go to a larger group to improve reliability and broaden authority. The final guidelines would present the strength of support and closeness of consensus for each statement, together with an appendix outlining the reasons given during the consensus process for surprising or controversial recommendations.

Acknowledgements

This study was funded by an MRC Clinician Scientist Fellowship for Rosalind Raine. We thank the study participants and steering committee, and Kirsten Larkin for providing administrative support and for assisting with the categorization of evidence of effectiveness.

References

1 Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: potential benefits, limitations and harms of clinical guidelines. *BMJ* 1999;**318**:527–30
2 Burgers JS, Grol R, Klazinga NS, Makela M, Zaat J, for the AGREE Collaboration. Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs. *Int J Qual Health Care* 2003;**15**:31–45
3 Delbecq A, Van de Ven A. A group process model for problem identification and program planning. *J Appl Behav Sci* 1971;**7**:467–92
4 Kastein MR, Jacobs M, Van der Hell RH, Luttick K, Touw-Otten FWMM. Delphi, the issue of reliability: a qualitative Delphi study in primary health care in the Netherlands. *Technol Forecast Soc Change* 1993;**44**:315–23
5 Washington DL, Bernstein SJ, Kahan JP, Leape LL, Kamberg CJ, Shekelle PG. Reliability of clinical guideline development using mail-only versus in-person expert panels. *Med Care* 2003;**41**:1374–81
6 Tobacman JK, Scott IU, Cyphert S, Zimmerman B. Reproducibility of measures of overuse of cataract surgery by three physician panels. *Med Care* 1999;**37**:937–45

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013

7 Escobar A, Quintana JM, Arostegui I, *et al*. Development of explicit criteria for total knee replacement. *Int J Technol Assess Health Care* 2003;**19**:57–70
8 Leape LL, Freshour MA, Yntema D, Hsiao W. Small-group judgment methods for determining resource-based relative values. *Med Care* 1992;**30**(Suppl):NS28–39
9 Wortman PM, Smyth JM, Langenbrunner JC, Yeaton WH. Consensus among experts and research synthesis: a comparison of methods. *Int J Technol Assess Health Care* 1998;**14**:109–22
10 Nicollier-Fahrni A, Vader JP, Froehlich F, Gonvers JJ, Burnand B. Development of appropriateness criteria for colonoscopy: comparison between a standardized expert panel and an evidence-based medicine approach. *Int J Qual Health Care* 2003;**15**:15–22
11 Eriksen BO, Almdahl SM, Hensrud A, *et al*. Assessing health benefit from hospitalization: agreement between expert panels. *Int J Technol Assess Health Care* 1996;**12**:126–35
12 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
13 Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. *N Engl J Med* 1998;**338**:1888–95
14 Hutchings A, Raine R, Sanderson C, Black N. An experimental study of determinants of the extent of disagreement within clinical guideline development groups. *Qual Saf Health Care* 2005;**14**:240–5
15 Raine R, Sanderson C, Hutchings A, Carter S, Larkin K, Black N. An experimental study of determinants of group judgments in clinical guideline development. *Lancet* 2004;**364**:429–37
16 Raine R, Haines A, Sensky T, Hutchings A, Larkin K, Black N. 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
17 Richardson FM. Peer review of medical care. *Med Care* 1972;**10**:29–39
18 Fitch K, Bernstein SJ, Aguilar MD, *et al*. *The RAND/UCLA Appropriateness Method User's Manual*. Santa Monica, Ca: RAND, 2001
19 Raine R, Carter S, Sensky T, Black N. General practitioners' perceptions of chronic fatigue syndrome and beliefs about its management, compared with irritable bowel syndrome: qualitative study. *BMJ* 2004;**328**:1354–7

Appendix

The random effects models to estimate differences in ratings and within-group agreement for each intervention were of the form $y_{ij} = \alpha + \beta x_{ij} + v_i + \varepsilon_{ij}$ for group i and scenario j. Semi-robust estimates of variance were calculated using the Huber/White/sandwich estimator.

Downloaded from hsr.sagepub.com at SAGE Publications on November 19, 2013