

An experimental study of determinants of group judgments in clinical guideline development

Rosalind Raine, Colin Sanderson, Andrew Hutchings, Simon Carter, Kirsten Larkin, Nick Black

Lancet 2004; 364: 429–37

See [Comment](#) page 392

Health Services Research Unit,
London School of Hygiene and
Tropical Medicine, Keppel
Street, London WC1E 7HT, UK
(R Raine PhD, C Sanderson PhD,
A Hutchings MSc, S Carter PhD,
K Larkin MSc, N Black MD)

Correspondence to:

Rosalind Raine
rosalind.raine@lshtm.ac.uk

Summary

Background Clinical guidelines for improving the quality of care are a familiar part of clinical practice. Formal consensus methods such as the nominal group technique are often used as part of guideline development, but little is known about factors that affect the statements produced by nominal groups, and on their consistency with the research evidence.

Methods Cognitive behavioural therapy, behavioural therapy, brief psychodynamic interpersonal therapy, and anti-depressants for irritable bowel syndrome, chronic fatigue syndrome, and chronic back pain were selected for study. 16 nominal groups in a factorial design allowed comparison of GP-only with mixed groups of GPs and specialists, provision of a literature review with no provision, and ratings made in the context of realistic or ideal levels of health-care resources. Participants rated appropriateness independently, and again after a facilitated meeting. Audiotapes of four group discussions were analysed.

Findings There was agreement with the research evidence for 51% of 192 scenarios. Agreement was more likely if the group was GP-only, if a literature review was provided, or if the evidence was in accordance with clinicians' beliefs. Assumptions about the level of resources available had no impact. Clinical and social cues had mixed effects, irrespective of the research evidence. Qualitative analysis showed the modifying effect of clinical experience and beliefs about research evidence.

Interpretation Guidelines cannot be based on data alone; judgment is unavoidable. The nominal group technique is a method of eliciting and aggregating judgments in a transparent and structured way. It can provide important information on levels of agreement between experts. However, conclusions can be at odds with the published literature. If they are, reasons need to be explicit.

Introduction

In many countries, there are clinical guidelines for disseminating good practice in medicine.^{1,2} Ideally, guidelines should be based on evidence from large, well conducted studies, but often such research does not exist³ and, where it does, how the results might be applied to particular patients can be unclear.¹ Also, guidelines may depend implicitly on interpretation of the literature, on judgments about value and risk,⁴ on the funding and organisation of health services,⁵ and, if public funding is involved, on policies about priorities and equity. The synthesis of the research evidence may be rigorous and transparent, but the judgments tend to be opaque.

Formal consensus development methods, often based on the nominal group technique, are widely used because, unlike informal methods such as committees, they offer structured, transparent, and replicable ways of synthesising individual judgments.⁶ In the UK, NICE and other professional bodies have used modified nominal group techniques, as have at least seven other countries.^{2,7,8}

In the modified nominal group technique, participants first express their views independently via a postal questionnaire. They then meet for review and discussion, after which they complete the questionnaire again privately, revising their views if they wish. The

practical application of this process has been far from uniform. A systematic review revealed a dearth of research into its workings⁹ and despite some subsequent studies,^{10–15} the key questions posed by the review have not yet been adequately answered. Our aim was to investigate the effect on the judgments produced and on the extent to which there was agreement with research evidence for: (1) three types of factor used to generate clinical scenarios provided in questionnaires—the clinical condition, the treatment, and clinical or social cues; (2) three ways in which nominal groups can differ—provision of a literature review or not, group composition, and background assumptions about the level of health-care resources available.

We also aimed to explore qualitatively the reasons behind the group judgments.

The other research priorities identified by the systematic review were to assess the reliability and representativeness of formal consensus techniques. Results of these investigations will be reported elsewhere.

Methods

Three conditions (chronic back pain, irritable bowel syndrome, and chronic fatigue syndrome) were selected because they fulfilled the following criteria:

(1) there was a mismatch between current clinical practice and research evidence; (2) care was provided by

Panel 1: Definition of interventions**Cognitive behavioural therapy**

This treatment is provided by cognitive behaviour therapists who aim to modify thoughts and beliefs thought relevant to the disease process with the expectation that emotional and behavioural changes will follow. It incorporates two elements: (1) a cognitive element, which includes an explanation of the cognitive model and identification of symptom-eliciting thoughts, feelings and behaviour; and (2) a behavioural element which includes behavioural experiments to test beliefs, the use of coping strategies, relaxation techniques, and biofeedback.

Some studies report the use of cognitive behavioural therapy, others report using cognitive therapy. For the purposes of the review, these interventions have been combined. There is no clear practical distinction between cognitive behavioural therapy and cognitive therapy, and the papers concerned do not give sufficient information to discriminate between the two interventions in terms of the therapeutic techniques used.

Behavioural therapies

This focuses on the modification of behaviour. Behaviour therapists seek to: (1) positively reinforce healthy behaviours. This approach emphasises the reinforcing role that social and environmental factors can play in the development and maintenance of functional somatic complaints. The goal is to identify and reinforce adaptive "well" behaviours (eg, exercising and talking about non-somatic topics) while reducing reinforcement for somatic behaviours (eg, excessive diagnostic testing or restricting mobility); or (2) modify physiological responses directly through using techniques such as biofeedback and relaxation training.

Brief psychodynamic interpersonal therapies

This approach addresses difficulties or problems in interpersonal relationships and the regulation of emotion. The nature of the patient's physical symptoms is explored by the psychotherapist, including the effect of the illness or condition on their lives, relationships and emotions. Dysfunctional patterns of interacting are identified and modified. The patient-therapist relationship is used as a model of change. Key techniques include: focusing on symptoms; using metaphor to explore emotional distress; working in the "here and now"; hypothesis generation; and the linkage of symptoms, feelings, and relationships to form an explanatory model.

Antidepressants:

These include tricyclics, selective serotonin reuptake inhibitors (SSRIs), and monoamine-oxidase inhibitors (MAOIs). The generic term antidepressant is used because the actual choice of antidepressant will depend on the patients' clinical presentation.

at least two groups of clinicians (general practitioners (GPs) and mental-health professionals); (3) these

Panel 2: Definition of realistic and ideal resource scenarios

By ideal, we mean 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,
- an inhouse general practice counsellor,

and

- the freedom to choose to whom you refer patients,
- short waiting times for good quality services,
- no financial barriers that limit the choice of treatment,
- timely and detailed feedback to the GP occurs.

By realistic we mean that patients can expect to wait for approximately:

- 6 weeks to see the inhouse 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),
- 6 weeks for an assessment at a pain clinic (followed by another wait of up to 6 months for treatment),

and

- access to clinical psychologists, liaison psychiatrists, and psychotherapists with expertise in managing patients with functional somatic symptoms is limited,
- services are organised on a geographical basis. This means that referral choices are limited,
- services are of variable quality in terms of their organisation and the range of clinical skills and experience that is available,
- no chronic fatigue syndrome clinic exists.

conditions are important problems; and (4) national guidelines for the conditions had not been published in the UK at the time the study materials were prepared in November, 2001.

Review of published work and development of the survey instrument

We conducted a systematic review of the evidence on the effectiveness of mental-health interventions in primary care patients who had any of chronic back pain, irritable bowel syndrome, or chronic fatigue syndrome. Four relevant interventions were identified: cognitive behavioural therapy, behavioural therapy, brief psychodynamic interpersonal therapy, and antidepressants (panel 1). Details of the search strategy, study selection, and synthesis have been reported elsewhere.¹⁶

A questionnaire was developed to elicit participants' judgement about appropriate treatment given the following clinical and social situations (ie, cues), identified

by GPs and psychiatrists: (1) co-existent depressive symptoms; (2) clinicians' perception that the patient believes that the condition has an organic cause; (3) insomnia in patients with chronic back pain; and (4) a financial motivation to return to work in patients with chronic fatigue syndrome. The questionnaire comprised 64 scenarios (eg, behavioural therapy for a patient with chronic fatigue syndrome who believes their condition has an organic cause). Participants were provided with definitions of the conditions¹⁷⁻¹⁹ and were instructed to rate treatment appropriateness on physical and psychological outcomes separately because the research evidence reported some differences in these respects. Every participant, therefore, made 128 ratings on Likert scales where 1=strong disagreement and 9=strong agreement.²⁰

Establishment of nominal groups

GPs and mental-health practitioners in England were identified from the Department of Health GP database, the Royal College of Psychiatrists liaison section database, and the British Association of Behavioural and Cognitive Psychotherapists database. For selection purposes we merged the two databases from the Royal College of Psychiatrists and the British Association for Behavioural and Cognitive Psychotherapists. Individuals were randomly selected for invitation to participate from the two lists with use of computer generated random numbers.²¹ The study had ethics approval from the London School of Hygiene and Tropical Medicine Ethics Committee.

Every combination of the three design factors—GPs only vs a mixed group; literature review provided vs not provided; decision making within the current resource availability in the UK NHS vs an ideal situation (panel 2)—was used in a factorial design (figure 1), producing eight types of group. To allow an assessment of reliability (to be reported elsewhere), each type was replicated once, resulting in 16 groups. All the groups that did not receive the literature review met before our review was published. With a target of 11 members per group, 14 were invited and between nine and 14 participated.

Data collection

Initial ratings were done by post. Each group then met for a facilitated meeting between February, 2002, and January, 2003. All 16 meetings were held at the same venue. The first group was facilitated by NB and all the others by RR. Facilitators followed a written protocol. At the meetings, every participant was given a new copy of the questionnaire, which included a reminder of their own initial ratings and those of the other members. Every scenario was discussed in turn, and reasons for any differences explored. The participants then privately rerated each scenario. Before leaving the meeting, participants completed an anonymised

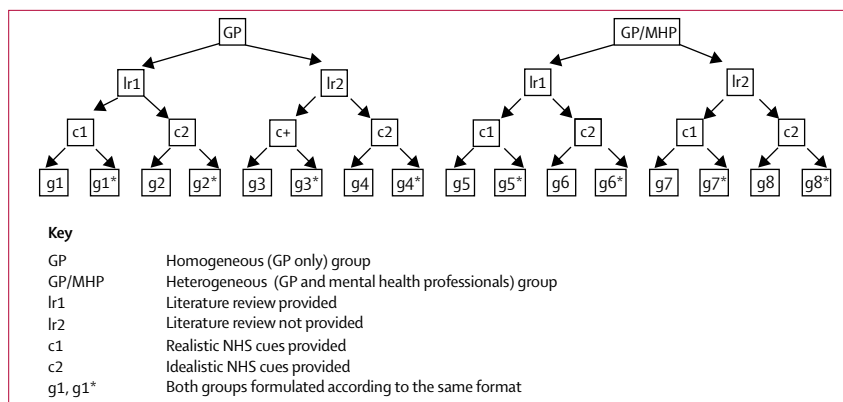


Figure 1: Study design: how three design factors were assessed in 16 nominal groups
MHP=mental-health professional.

questionnaire covering their demographic characteristics.

The meetings were audiotaped and RR wrote field notes describing group processes and the key issues discussed immediately after each meeting. The audiotapes were then transcribed verbatim (KL).

Analysis

To test the representativeness of participants' views, we compared the initial ratings of four of the nominal groups with those produced by two larger groups of randomly selected GPs and mental-health practitioners. Two nominal groups comprising ten and nine GPs were compared with a group of 85 GPs, and two nominal groups of seven GPs and six mental-health practitioners, and of five GPs and seven mental-health practitioners respectively, were compared with a group of 43 GPs and 41 mental-health practitioners. Agreement of the groups' median scores was assessed with a weighted kappa statistic (κ_w). All groups compared were provided with a literature review and assumed a realistic level of resources.

Comparison of groups' final ratings with research evidence

We calculated a median rating for each group for each of the 12 condition/treatment combinations (eg, behavioural therapy in chronic fatigue syndrome). On the basis of results of the systematic review, two of the authors (RR and KEL) independently categorised the research evidence into one of three groups: good evidence of no benefit; evidence lacking or equivocal; and good evidence of benefit. There was no research evidence for varying treatment according to cue except for depression, which influenced the effectiveness of treatment for irritable bowel syndrome and chronic fatigue syndrome. For comparison with the groups' ratings, research evidence was deemed consistent with group ratings of 1.0–3.67 if the intervention was inappropriate, 3.68–6.35 (intervention equivocal or uncertain), and 6.36–9.0 (intervention appropriate).

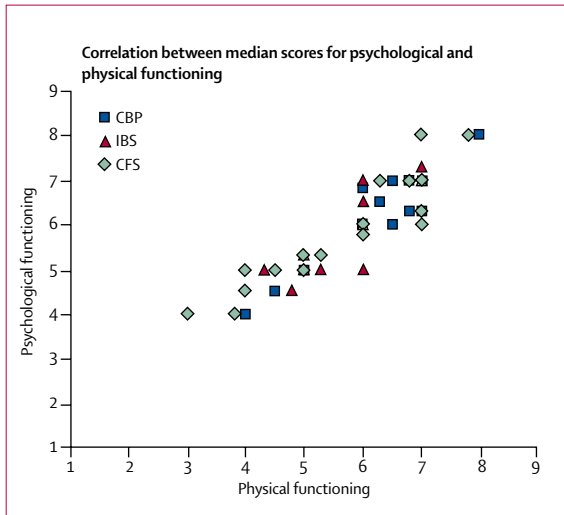


Figure 2: Comparison of ratings for physical and for psychological outcomes

Effects of clinical and social cues

All of the 16 groups produced a median for each scenario. To summarise across groups, we calculated medians of group medians for every scenario. For all

condition/treatment combinations we used the Sign test for paired differences between group median ratings with and without the cue. This test does not assume equal intervals across the Likert scales.

Effects of design factors

For each group with a given factor (eg, supplied with a literature review) we calculated a median rating across all cues and participants for every condition/treatment combination. This median rating was compared both with the corresponding median for the groups without the factor and with the research evidence.

We also used ANOVA to assess whether the three design factors were associated with differences in ratings between the groups. Although this method involves calculation of means of medians, it provided a comprehensive approach to the analysis. The treatment, condition, and level of evidence from the research literature were included as within-group factors and interactions with the design factors were tested. A Bonferroni adjustment ($p < 0.004$) was used for all interactions. The effect sizes and their 95% CI were calculated from a regression model with each group treated as a random effect.

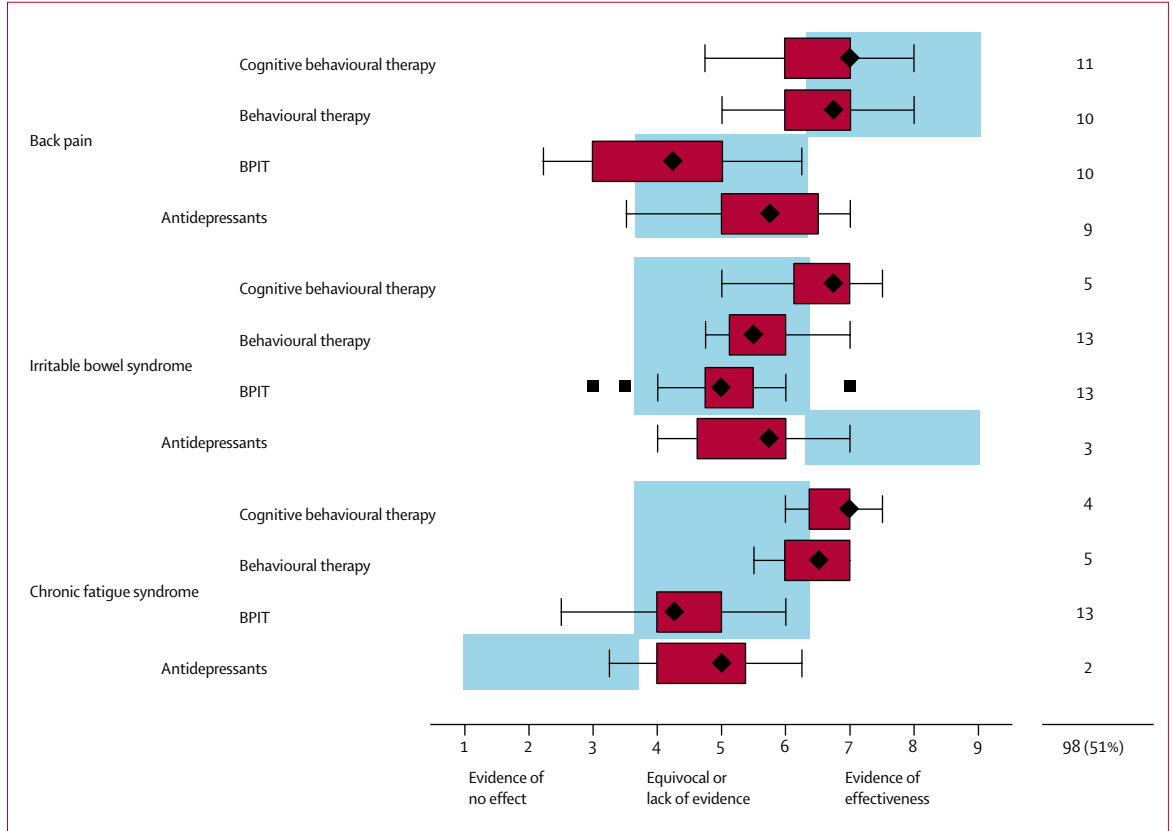


Figure 3: Distribution of the 16 group ratings

BPIT=brief psychodynamic interpersonal therapy. Background shading shows our judgments about the research evidence. Diamonds show median and box shows interquartile range; whisker show the range of ratings within 1.5xIQR. Squares show outliers. Number in right column shows number of groups in agreement with the evidence.

	Treatments			
	CBT	BT	BPIT	Antidepressants
Chronic back pain				
Research evidence	6.36–9.00	6.36–9.00	3.68–6.35	3.68–6.35
Scenario				
Symptoms of depression	7.0	6.8	5.0	8.0
No depression symptoms	6.8	7.0	4.0	5.0
p for difference*	0.0005	1.00	0.001	<0.0001
Problem believed organic	7.0	6.5	4.0	5.0
Would try anything	7.0	7.0	5.0	6.0
p for difference*	0.002	0.001	0.04	0.001
Pain insomnia	7.0	6.3	4.0	6.5
No insomnia	7.0	6.0	4.5	5.0
p for difference*	0.29	1.00	0.45	0.0001
Irritable bowel syndrome				
Research evidence	3.68–6.35	3.68–6.35	3.68–6.35	6.36–9.00
Scenario				
Symptoms of depression	7.0	5.0	5.0	7.0
No depression symptoms	6.0	5.3	5.0	4.8
p for difference*	0.18	1.00	0.51	0.0001
Problem believed organic	6.0	6.0	4.3	5.0
Exacerbated by stress	7.0	7.0	6.0	6.0
p for difference*	0.04	0.001	0.0001	0.02
Chronic fatigue syndrome				
Research evidence	3.68–6.35	3.68–6.35	3.68–6.35	1.00–3.67
Scenario				
Return to work	7.0	7.0	4.0	5.0
No economic contribution	6.3	6.0	4.5	5.0
p for difference*	0.45	0.01	0.34	0.38
Symptoms of depression	7.0	6.0	5.0	7.0
No depression symptoms	7.0	7.0	4.0	3.0
p for difference*	0.18	0.002	0.34	<0.0001
Problem believed organic	6.8	6.0	3.8	4.0
Would try anything	7.8	7.0	5.0	5.3
p for difference*	0.001	<0.0005	0.0002	0.01

*Non parametric paired comparison using the Sign test. The Sign test can give significant results despite equal medians because it tests the direction of differences in ratings, not equality of medians. CBT=cognitive behavioural therapy. BT=behavioural therapy. BPIT=brief psychodynamic interpersonal therapy.

Table 1: Comparison of group median ratings with research evidence for different clinical scenarios

Responses to research evidence

On completion of the quantitative analysis, we used qualitative methods to explore why groups' ratings might diverge from the research evidence. Two of the investigators (RR and SC) independently scrutinised entire transcripts of group discussions and drew up a preliminary list of themes. The themes were then discussed by RR and SC. A variant of grounded theory was applied whereby provisional themes were first identified with use of the group's own concepts and these were applied to later transcripts to allow the emergence of an analytical theory suited to the context.²² This process was continued until no new themes emerged, which occurred after the analysis of four transcripts. Careful attention was given to so called deviant cases that did not fit in with the concepts and emerging themes.²³ The themes that emerged from the analysis of the first four transcripts were checked against field notes taken in the other 12 groups and presented to the study's Advisory Committee for comment and endorsement.

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, the writing of the report, or the decision to submit the report for publication.

Results

There were 177 participants in the 16 groups, of whom 76% were GPs and 24% mental-health professionals. Mean age was 47 years, most were men (62%) and white (84%).

The relation between ratings for physical and for psychological outcomes for each scenario across the 16 groups was assessed by plotting their medians (figure 2). In view of the close agreement, subsequent analyses used the physical outcome ratings only.

Initial ratings produced by the GP-only nominal groups showed moderate agreement ($\kappa_w=0.53$, 95% CI 0.43–0.63) with those produced by the larger group of GPs. The ratings produced by the mixed nominal groups showed substantial agreement (0.76, 0.65–0.88) with those produced by a larger mixed group.

Median values for each group are shown as box-and-whisker plots in figure 3. There was evidence of effectiveness or ineffectiveness for only four of the 12 condition/treatment combinations. Of the 192 group median ratings (12 condition/treatment combinations, 16 groups), 98 (51%) agreed with the evidence. The amount of agreement varied according to the condition considered: 63% for chronic back pain, 59% for irritable bowel syndrome, and 38% for chronic fatigue syndrome. Groups were consistently equivocal about the use of antidepressants despite unequivocal research evidence in irritable bowel syndrome and chronic fatigue syndrome.

Some clinical and social cues had striking effects on group ratings (table 1). The presence of depressive symptoms made groups more likely to recommend antidepressants for all three conditions, despite research

	Group composition*		Difference (95% CI)	p
	Mixed	GPs only		
CBT	6.85	6.50	0.35 (–0.14 to 0.85)	<0.0001
BT	6.33	6.19	0.15 (–0.35 to 0.64)	
BPIT	3.93	5.07	–1.15 (–1.64 to –0.66)	
Antidepressants	5.06	5.64	–0.57 (–1.07 to –0.08)	
Literature review	Provided	Not provided		
Evidence of effectiveness	6.81	5.78	1.03 (0.54 to 1.52)	<0.0001
Equivocal/lack of evidence	5.59	5.58	0.01 (–0.36 to 0.37)	
Evidence of no effect	4.44	5.13	–0.69 (–1.46 to 0.09)	
Context	Realistic	Idealistic	Context	
	5.60	5.79	–0.19 (–0.52 to 0.14)	

CBT=cognitive behavioural therapy. BT=behavioural therapy. BPIT=brief psychodynamic interpersonal therapy. *Data are mean of median ratings; positive ratings indicate stronger agreement of effectiveness.

Table 2: Effect of group level factors on differences in median ratings

	Rating based on research evidence*	Group composition			
		GP only (n=8)		Mixed (n=8)	
		Median rating	Groups agreeing with evidence	Median rating	Groups agreeing with evidence
Chronic back pain					
CBT	6.36-9.00	7	5	7	6
BT	6.36-9.00	6.75	5	6.75	5
BPIT	3.68-6.35	5	7	3†	3
Antidepressants	3.68-6.35	6	4	5.5	5
Irritable bowel syndrome					
CBT	3.68-6.35	6.75†	3	6.75†	2
BT	3.68-6.35	5.5	7	5.5	6
BPIT	3.68-6.35	5	7	5	6
Antidepressants	6.36-9.00	6†	2	5.25†	1
Chronic fatigue syndrome					
CBT	3.68-6.35	6.5†	4	7†	0
BT	3.68-6.35	6.5†	3	6.88†	2
BPIT	3.68-6.35	5	8	4	5
Antidepressants	1.00-3.67	5.13†	0	4.13†	2
Total of group ratings agreeing with evidence (%)			55 (57.3%)		43 (44.8%)

*If evidence only in secondary care, counts as equivocal. †Instances where the observed rating differs from the expected rating.

Table 3: Median ratings for homogeneous compared with heterogeneous nominal groups

	Rating based on research evidence*	Context			
		Realistic (n=8)		Ideal (n=8)	
		Median rating	Groups agreeing with evidence	Median rating	Groups agreeing with evidence
Chronic back pain					
CBT	6.36-9.00	7	5	7	6
BT	6.36-9.00	7	6	6.38	4
BPIT	3.68-6.35	3.5†	4	4.75	6
Antidepressants	3.68-6.35	5.75	6	6	3
Irritable bowel syndrome					
CBT	3.68-6.35	6.13	5	7†	0
BT	3.68-6.35	5.38	6	5.75	7
BPIT	3.68-6.35	5.13	5	5	8
Antidepressants	6.36-9.00	5.75†	1	5.38†	2
Chronic fatigue syndrome					
CBT	3.68-6.35	6.88†	3	7†	1
BT	3.68-6.35	6.5†	3	6.5†	2
BPIT	3.68-6.35	4	5	4.75	8
Antidepressants	1.00-3.67	5†	0	4.75†	2
Total of group ratings in agreement with evidence (%)			49 (51.0%)		49 (51.0%)

*If evidence only in secondary care, counts as equivocal. †Instances where the observed rating differs from the expected rating.

Table 5: Median ratings for nominal groups working with current versus ideal levels of resources

evidence that the effectiveness of antidepressants is unrelated to depression in chronic fatigue syndrome and irritable bowel syndrome. Clinicians' perceptions of patients' beliefs about the basis of their condition also affected ratings. Patients with chronic fatigue syndrome

who believed their condition was organic were thought to be much less likely to benefit from any of the psychological treatments. This belief was also thought to limit the effectiveness of behavioural therapy and brief psychodynamic interpersonal therapy in irritable bowel syndrome, and of cognitive behavioural therapy, behavioural therapy, and antidepressants in patients with chronic back pain.

	Rating based on research evidence*	Literature review			
		Supplied (n=8)		Not supplied (n=8)	
		Median rating	Groups agreeing with evidence	Median rating	Groups agreeing with evidence
Chronic back pain					
CBT	6.36-9.00	7	8	6†	3
BT	6.36-9.00	7	7	6†	3
BPIT	3.68-6.35	4.5	5	4.25	5
Antidepressants	3.68-6.35	5.25	5	6.38†	4
Irritable bowel syndrome					
CBT	3.68-6.35	6.75†	3	6.75†	2
BT	3.68-6.35	5.5	8	6	5
BPIT	3.68-6.35	5.25	6	5	7
Antidepressants	6.36-9.00	6†	3	4.75†	0
Chronic fatigue syndrome					
CBT	3.68-6.35	7†	2	7†	2
BT	3.68-6.35	6.5†	2	6.5†	3
BPIT	3.68-6.35	4.75	7	4	6
Antidepressants	1.00-3.67	4.38†	2	5.13†	0
Total of group ratings in agreement with evidence (%)			58 (60.4%)		40 (41.7%)

*If evidence only in secondary care, counts as equivocal. †Instances where the observed rating differs from the expected rating.

Table 4: Median ratings for nominal groups provided with research evidence compared with those without research evidence

Economic motivation was not seen as making an important difference to the appropriateness of any treatment for chronic fatigue syndrome, and pain-induced insomnia did not affect views of the appropriateness of any psychological treatment in patients with chronic back pain, although antidepressants were believed to be much more appropriate for those with insomnia than those without it.

Table 2 shows the significant interactions in the ANOVA. Mixed groups tended to rate brief psychodynamic interpersonal therapy and antidepressants lower than did the GP-only groups. 45% of their ratings agreed with the evidence, compared with 57% for GP-only groups (table 3).

Groups with a literature review agreed with the evidence more often (60% of the 192 group median ratings) than did the groups without a review (42%, table 4), producing higher ratings where there was evidence of effectiveness and lower ratings where there was evidence of ineffectiveness (table 2). We did not note any evidence that resource context had any effect on group ratings (table 2) or on agreement with the evidence (table 5).

Panel 3: Factors affecting extent to which groups' ratings agree with research evidence

Reasons for discounting evidence

Weak or irrelevant evidence

"I thought that this evidence was basically unhelpful and omitted a huge area of relevance" (GP 14)

"Any study will show an effect, even if it's a very slight effect and that may not be all that clinically helpful" (GP 114)

"The fact that there are no studies doesn't mean it's no good" (MHP 41)

Clinical experience

"Anybody locally who goes to the pain clinic routinely gets referred for CBT [cognitive behavioural therapy] with impressive results" (GP 21)

"On the evidence you should be saying it is not a good treatment option. But experience was that some people . . . got better with it . . . and it is what you yourself would want. And so everything is moved up . . . a notch" (GP 13)

"You look for evidence to try and guide you, but you are not really sure that the patients who are in the trials are actually the same as your patients and so you tend to fall back on your own beliefs or your own experience" (GP 13)

Patient preference

"If the evidence is not strong, you might listen to what they (patients) want" (GP 14)

Treatment availability

"It is a mixture of both having experts available in our area in addition to the evidence" (Mental-health practitioner 39)

Unwillingness to do nothing

"I want to be doing something, better than nothing" (GP 44)

Reasons for relying on evidence

Evidence perceived as credible

"I scored the way I did because the evidence you gave us suggested that that was the way to go" (MHP 39)

Coincidence of lack of evidence with lack of knowledge

"That was a combination of not particularly knowing enough about it to feel confident to recommend it to anybody and seeing that there was no evidence" (GP 18)

Groups made recommendations that diverged from research findings for five main reasons: weak or irrelevant evidence, clinical experience, patient preference, treatment availability, and reluctance to do nothing (panel 3). Participants stated that they were more likely to accept evidence if it supported their current practice.

Evidence that effectiveness varied with condition was often ignored. Participants tended to believe in cognitive behavioural therapy as a universal remedy that "is brilliant for most things" (GP 10). Likewise, some participants favoured antidepressants simply because of their availability ("This is something GPs can do immediately. Referring to other services is a six months

waiting list" [GP 19]), their cost, and a desire to be seen to be doing something. GPs who were against the use of antidepressants cited side-effects, negative connotations ("People think that you are saying that they are making it up" [GP 10]) and unhelpfully medicalising their illness ("It buys into their medical model of their illness" [GP 18]).

The effect of cues was much the same across conditions, irrespective of the evidence. Benefits of antidepressants for patients with coexistent depression were believed to be self evident ("What do you treat depression with? Antidepressants" [GP 9]). Patients who believed their conditions had an organic cause were not seen as suitable for psychological therapies because of the potential for resistance ("like bashing your head against a lump of concrete" [mental-health practitioner 1]). Antidepressants for insomnia were favoured because of their analgesic and sedative properties. However, participants felt uncomfortable about passing judgment on the effect of economic motivation.

Participants were influenced by evidence where it was perceived to be credible. They also made judgments based on an absence of evidence when this coincided with their own lack of knowledge about a treatment, such as brief psychodynamic interpersonal therapy.

Dialogue between professional groups tended to increase divergence from the evidence. This was either because of knowledge transfer between professions ("Not being a doctor I hadn't thought about the side-effects of antidepressants. I would change my rating", [mental-health practitioner 4]) or the wide range of experiences in mixed groups that included events where evidence was seen not to apply.

We noted that resource context had little effect on ratings but with ideal resource assumptions, participants admitted to "carry(ing) over our own prejudices and habits" (mental-health practitioner 42).

Discussion

A formal consensus development method produced judgments that were consistent with our assessments of the research evidence in about half the scenarios considered. The extent of concordance varied between the conditions and treatments studied. Concordance was more likely if a literature review was provided and if this evidence supported clinicians' experiences and beliefs. If clinical experience and beliefs were not consistent with research evidence, then the experience and beliefs seemed to take precedence. Also, we postulate a so-called halo effect, in which evidence that a treatment is effective for one condition leads to unduly favourable opinions of its use for others. Agreement with the research evidence did not depend on the clarity of the research message; concordance was just as likely when the literature was equivocal. Group composition affected both the ratings and concordance with the evidence. The resource context had little effect.

Because of the tendency for specialists to over-state the effectiveness of their specialist intervention^{14,15} our nominal groups included generalists as well as specialists. Both had experience of managing patients with the three conditions in situations where guidelines would apply.

Treatment guidelines should reflect not only on technical judgments about effectiveness but also on the opportunity costs of other interventions. The idea that evidence is translated through a filter of social and organisational factors has been addressed elsewhere.^{4,24} Ideally, guidelines would be developed with use of expert groups to estimate treatment effectiveness, patients and the public to provide information about the value of different outcomes, and economists to estimate costs. We made no attempt to separate fact and value, and supplied no data on costs. That the resource cue had little effect on participants' judgments suggests that the groups concerned themselves chiefly with effectiveness.

There was a potential conflict between what participants thought they should score, given the status of the evidence, and their own practice and beliefs. This factor is likely to affect not only participants' views of research evidence but also the likelihood of guidelines being implemented. There was also potential for participants to focus on different psychological outcomes in irritable bowel syndrome and chronic back pain, since psychological symptoms were not specified in the definitions of the conditions that were provided to participants. This definition was in contrast to those provided for chronic fatigue syndrome, which included physical and psychological symptoms. However, this difference did not have much effect as both physical and psychological outcomes for all three conditions were widely discussed.

These findings may not be generalisable to conditions with fewer psychosocial determinants and a clearer pathogenesis. Furthermore, concordance between group judgments and the research evidence is conditional on the interpretation of the research evidence. The research evidence was assessed independently by two reviewers, but their assessments were necessarily subjective.¹⁶

Previous studies that have looked at the effect of design factors on group judgments have focused on the ratings produced by individual participants rather than the entire group because of the small numbers of groups involved.^{13,14,25,26} This is the first study that has assessed the effect of design factors on the judgments produced by the group, with the group as the unit of analysis.²⁷

In our study, agreement between group judgments and research evidence was slightly less than that achieved in a Swiss study on clinical indications for colonoscopy.¹⁰ It would be surprising if agreement did not vary with topic and it may vary between health-care systems. We noted that resource assumptions did not

affect ratings, but this may be because clinicians find it difficult to imagine working with unlimited resources. Another study that compared expert consensus with the research evidence reported mixed results.¹¹ Our findings support the idea that evidence is used to confirm pre-existing opinions rather than change them.²⁸

Other studies that have assessed the effect of group composition on consensus decisions have focused on surgical conditions.¹²⁻¹⁵ By contrast with our findings, the investigators noted that judgments often reflected specialist interests, but they did not investigate whose judgments were most closely aligned with the evidence. Finally, one study noted that consultants were the most active contributors to a guideline development panel, followed by GPs, and then professions allied to medicine.²⁹ We did not attempt to quantify the contributions made by the different health professionals in the sixteen groups, but every professional group certainly made a contribution. It may be that nominal groups comprising health-care professionals who are already familiar with each other relate in a different way to members of panels who have not met before, and whose status is therefore less clear, as in our study.

Guidelines cannot be deduced from research evidence alone. Statements about what ought to be done in particular circumstances necessarily depend on interpretation of the evidence and on clinicians' experience, beliefs, and values. It could be argued that transparency about the process of synthesising evidence and beliefs is even more important than transparency about the evidence itself, which should be in the public domain anyway. However, how best to do this remains unclear.

The nominal group technique is a method of eliciting and aggregating individual syntheses of evidence and belief in a transparent and structured way, and it provides important information on levels of agreement about treatment. However, judgments can be at odds with the published literature. If they are, the reasons for this need to be made explicit.

Contributors

R Raine, N Black, and C Sanderson designed the study; R Raine and K Larkin collected the data; R Raine, N Black, C Sanderson, and A Hutchings planned statistical analyses; R Raine, C Sanderson, and A Hutchings did statistical analysis; qualitative analyses were planned and done by R Raine and S Carter. All investigators reviewed and approved the final manuscript.

Conflict of interest statement

None declared.

Acknowledgments

We thank the GPs and mental-health specialists who participated, Andy Haines and Jan van der Meulen for comments, and the Medical Research Council for funding a Clinician Scientist Fellowship for RR.

References

- 1 Woolf H, Grol R, Hutchinson A, Eccles M, Grimshaw J. 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 programmes. *Int J Qual Health Care* 2003; **15**: 31–45.
- 3 Chassin M. How do we decide whether an investigation or procedure is appropriate? In: Hopkins A, ed. *Appropriate investigation and treatment in clinical practice*. London: Royal College of Physicians, 1989; 21–29.
- 4 Fitzgerald L, Ferlie E, Hawkins C. Innovation in healthcare: how does credible evidence influence professionals? *Health Soc Care Community* 2003; **11**: 219–28.
- 5 Mann T. *Clinical guidelines: using clinical guidelines to improve patient care in the NHS*. London: Department of Health, 1996.
- 6 Delbecq A, van de Ven A. A group process model for problem identification and programme planning. *J Appl Behav Sci* 1971; **7**: 467–92.
- 7 Glasier A, Brechin S, Raine R, Penney G. A consensus process to adapt the World Health Organisation selected practice recommendations for UK use. *Contraception* 2003; **68**: 327–33.
- 8 NICE (UK) National Collaborating Centre for Acute Care (UK). *Preoperative tests: the use of routine preoperative tests for elective surgery*. London: National Institute of Clinical Excellence, 2003.
- 9 Murphy M, Black N, Lamping D, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assessment* 1998; **2**: 1–88.
- 10 Nicollier-Fahrni A, Vader J, Froehlich F, Convers J, 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 Wortman P, Smyth J, Langenbrunner J, Yeaton W. Consensus among experts and research synthesis: a comparison of methods. *Int J Technol Assess Health Care* 1998; **14**: 109–22.
- 12 Bernstein S, Lazaro P, Fitch K, Aguilar M, Kahan J. Effect of specialty and nationality on panel judgement of the appropriateness of coronary revascularisation: a pilot study. *Med Care* 2001; **39**: 513–20.
- 13 Landrum M, McNeil B, Silva L, Normand S-L. Understanding variability in physician ratings of the appropriateness of coronary angiography after acute myocardial infarction. *J Clin Epidemiol* 1999; **52**: 309–19.
- 14 Fitch K, Lazaro P, Aguilar M, Martin Y, Bernstein S. Physician recommendations for coronary revascularization: variations by clinical specialty. *Eur J Public Health* 1999; **9**: 181–87.
- 15 Herrin J, Etchason J, Kahan J, Brook R, Ballard D. Effect of panel composition on physician ratings of appropriateness of abdominal aortic aneurysm surgery: elucidating differences between multi-specialty panel results and specialty society recommendations. *Health Policy* 1997; **42**: 67–81.
- 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**: 1082–86.
- 17 Reid S, Chalder T, Cleare A, Hotopf M, Wessely S. Extracts from clinical evidence: chronic fatigue syndrome. *BMJ* 2000; **320**: 292–96.
- 18 Manning A, Thompson W, Heaton K, Morris A. Towards a positive diagnosis of the irritable bowel. *BMJ* 1978; **2**: 653–54.
- 19 Vlaeyen J, Haazen I, Schuerman J, Kole-Snijders A, van Eek H. Behavioural rehabilitation of chronic low back pain: comparison of an operant treatment, an operant-cognitive treatment and an operant-respondent treatment. *Br J Clin Psychol* 1995; **34**: 95–118.
- 20 Streiner D, Norman G. *Health measurement scales: a practical guide to their development and use*, 2nd edn. Oxford: Oxford University Press, 1995.
- 21 SPSS. *SPSS base 10.0 for windows users guide*. Chicago: SPSS, 1999.
- 22 Green J. Grounded theory and constant comparative method. *BMJ* 1998; **316**: 1064–65.
- 23 Green J. Commentary: generalisability and validity in qualitative research. *BMJ* 1999; **319**: 421.
- 24 Ferlie E, Fitzgerald L, Wood M. Getting evidence into clinical practice: an organisational behaviour perspective. *J Health Serv Res Policy* 2000; **5**: 96–102.
- 25 Ayanian J, Landrum M, Normand S-L, Guadagnoli E, McNeil B. Rating the appropriateness of coronary angiography: do practicing physicians agree with an expert panel and with each other? *N Engl J Med* 1998; **338**: 1896–904.
- 26 Campbell S, Hann M, Roland M, Quayle J-A, Shekelle P. The effect of panel membership and feedback on ratings in a two-round Delphi survey. *Med Care* 1999; **37**: 964–68.
- 27 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**: 44–48.
- 28 Sauerland S, Neugebauer E. Consensus conferences must include a systematic search and categorization of the evidence. *Surg Endosc* 2000; **14**: 908–10.
- 29 Pagliari C, Grimshaw J. Impact of group structure and process on multi-disciplinary evidence-based guideline development: an observed study. *J Eval Clin Pract* 2002; **8**: 145–53.