ESSAYS

Seeking consensus by formal methods: a health warning

Carol Tan[1] Tom Treasure[1] John Browne[2] Martin Utley[3] Christopher W H Davies[4] Harry Hemingway[5]

J R Soc Med 2007;**100**:10–14

There are compelling arguments for the creation and use of guidelines to steer clinical practice: the range of interventions and therapies on offer is large and ever increasing; more effective treatments should be preferred over the less effective; costs are often high and resources are limited so funds should be focused where they will do the most good for the most patients. Some doctors are antipathetic to guidelines—which they decry as 'cookbook medicine'—but there is now general acceptance that we cannot afford a free-for-all in clinical practice. The authors believe that there should be equitable access to health care. Expenditure on relatively ineffective treatments, and treatment at high cost of those who can pay to the detriment of those who cannot, run counter to this principle.

In order to adjudicate on what should and should not be done, we need transparent processes. There are several systems advocated for the grading of evidence,[1] with randomized controlled trials gaining an A grade in an A, B, C grading. There are many aspects of care which rely on lower levels of evidence but when the cause and effect relationship between treatment and outcome is clear, the beneficial effect is large, and cost is supportable, then a strong recommendation (grade 1 of 2) can still be made. Hip replacement, cataract surgery and valve replacement for aortic stenosis are all in this league. Trials of effectiveness are not justified at this stage, and the 'sky diving without a parachute' analogy[2] can be invoked. They earn 1C recommendation: that is, a strong recommendation with poor quality evidence.[1]

There are many instances where evidence of benefit is much less clear or the effect size is marginal. Where high level evidence runs out, increasing use is made of expert panels. In this paper we illustrate pitfalls in the expert panel process as used for ratings of appropriateness of interventions for the very common clinical problem of malignant pleural effusion,[3] and point to areas for future development.

OUR CLINICAL PROBLEM—MALIGNANT PLEURAL EFFUSION

Pleural effusion is a manifestation of disseminated cancer estimated to affect 100 000 patients per year in the United States, often in the last few months of life.[4] It is a cause of debilitating but relievable breathlessness. Drainage of the effusion may give dramatic temporary relief and, if so, better breathing can be maintained by chemically induced pleurodesis.[5] Pleurodesis can be performed under general anaesthetic by a thoracic surgeon, or at the bedside. Its timely use in appropriate cases can give great relief but clinicians have widely differing knowledge and experience in the treatment of malignant pleural effusion, leading to variation in practice. The treatment, poorly implemented in an unsuitable case, is worse than useless. A Cochrane systematic review[6] and our own more broadly scoped systematic review[7] revealed the paucity of high quality evidence for most considerations, so how are we to formulate guidelines?

WHAT WE KNOW FROM RCTS

The published randomized trials recruited patients with large effusions, a non-trapped lung and a prognosis of at least one month. In such patients, the trial evidence supports pleurodesis as giving symptomatic benefit. In practice we see patients who would have been ideal candidates for the procedure some months ago but instead have been put through several cycles of aspiration and recurrence before it is considered. The ideal opportunity may have been lost. On other occasions a dying patient is referred for a surgical intervention when all else has failed and we believe no useful relief can be obtained. So pleurodesis may range from highly appropriate through unavailing and futile to positively detrimental—but how is this decided? An ad hoc clinical decision is made by whoever sees the patient but it would be better if written advice in the form of a guideline were available to aid appropriate decision making. Previously such guidance was derived from a meeting of 'the great and the good' in a given field who were invited to pool their experience. This method is referred to irreverently as the GOBSAT method

[1]Thoracic Unit, Guy's Hospital, St Thomas' Street, London SE1 9RT, UK
[2]Health Services Research Unit, London School of Hygiene and Tropical Medicine and Clinical Effectiveness Unit, Royal College of Surgeons of England, Keppel Street, London WC1E 7HT, UK
[3]Clinical Operational Research Unit, UCL, London WC1E 6BT, UK
[4]Royal Berkshire Hospital, Reading, Berkshire RG1 5AN, UK
[5]Department of Epidemiology, UCL, London WC1E 6BT, UK
Correspondence to: Tom Treasure
Email: Tom.Treasure@gmail.com

Talc pleurodesis for malignant pleural effusion

CHAPTER 1 Life expectancy <3 months	Appropriateness of VATS & talc plurodesis		Appropriateness of bedside talc slurry	
	Pleural thickening present	Pleural thickening absent	Pleural thickening present	Pleural thickening absent
A. Dyspnoea score 1 (breathless only with strenuous exercise				
1. Symptomatic relief following aspiration or drainage				
Trapped lung present	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung absent	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung not definable	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
2. No change in symptoms after aspiration or drainage				
Trapped lung present	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung absent	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung not definable	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
3. Symptoms worse following aspiration or drainage				
Trapped lung present	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung absent	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung not definable	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
4. Aspiration or drainage not done				
Trapped lung present	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung absent	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung not definable	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
B. Dyspnoea score 2 (breathless when hurrying on level or up slight hill)				
1. Symptomatic relief following aspiration or drainage				
Trapped lung present	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung absent	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9
Trapped lung not definable	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9

VATS: Video Assisted Thoracic Surgery

Figure 1

(good old boys sat around a table). Uncertainties and disagreements are resolved in various non-transparent ways.

In the case of malignant pleural effusion, the clinician authors, aware of the variability in practice, took the problem to experts in health service research to seek the best way towards statements of appropriateness. This paper reflects on our shared view of the process.

AVAILABLE METHODS

We wanted a method to formalize the process of constructing a decision tree.

Amongst others, the US Agency for Health Care Research and Quality (AHCRQ) and the UK National Institute for Clinical Excellence (NICE) have sought to address these limitations by using formal methods of expert panel judgements. One favoured process is the RAND Appropriateness Model (RAM).[8] Following the method, we convened a panel comprising three respiratory physicians, three thoracic surgeons and two oncologists and gave them a summary of the available evidence in the form of a systematic review.[7] The scope is designed to be comprehensive—unlike in randomized trials, all patients seen in clinical practice are included in the frame. The various factors that might be considered are used to construct a matrix (Figure 1), creating a number of permutations which should be sufficient, but no more than required, to create descriptive subsets of patients sufficiently homogeneous that the same rating would apply to all. Informed by the factors considered in the trials,[7] we identified five clinical attributes that might reasonably be taken into account in deciding the appropriateness of pleurodesis for an individual patient.

The panel rated on paper the appropriateness of surgical pleurodesis performed either by video assisted thoracic surgery (VATS) or at the bedside through a chest drain, for numerous hypothetical scenarios. Appropriateness was rated on a scale from 1 (highly inappropriate) to 9 (highly appropriate). Subsequently, panelists attended a one-day meeting to discuss and review their judgments with a facilitator experienced in the method.[9,10]

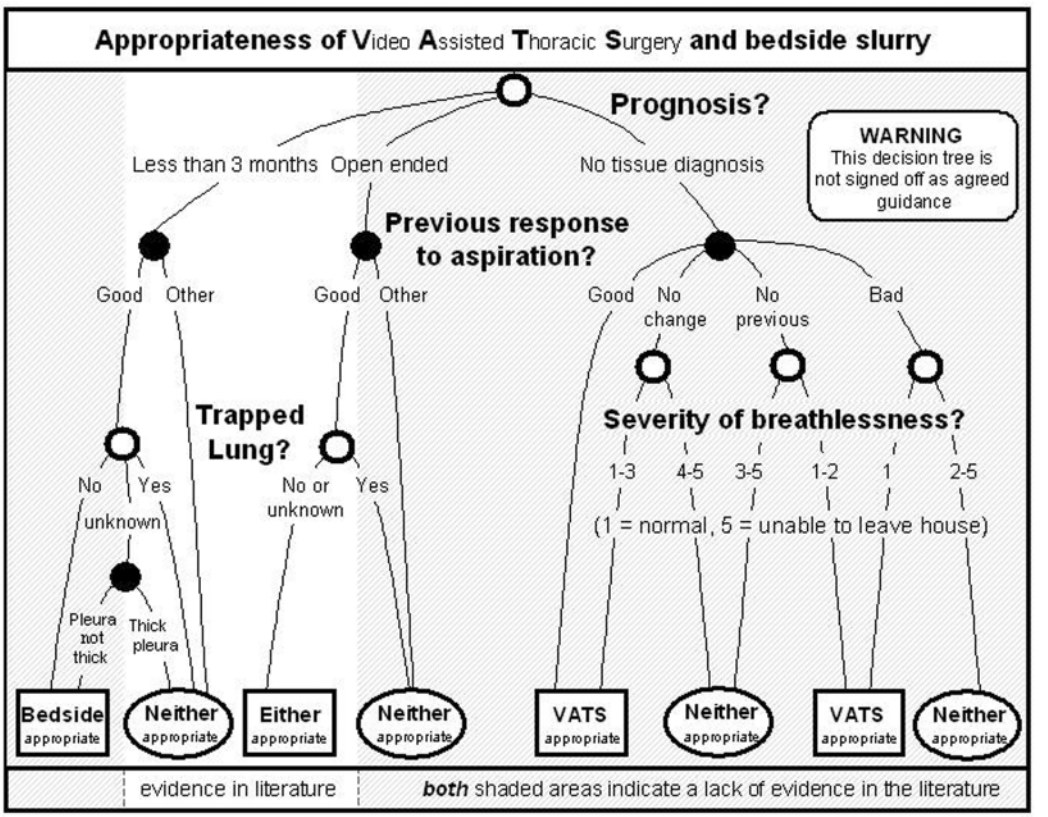

Figure 2

RESULTS

The full results of the expert panel are published elsewhere[3] and are summarized in the decision tree (Figure 2), but have to be presented with a 'health warning'. Just where the method should have helped—in the negotiable and opinion based areas where there are no data—it appeared to let us down.

One method of summarizing this process is to construct a decision chart. This involved building a model of how clinical judgements were influenced by patient attributes. We successfully reproduced the panel ratings of appropriateness and found a clear hierarchy amongst the clinical descriptors.

Two results we did not trust

Where life expectancy was less than three months, surgical pleurodesis was never deemed appropriate by the panel, and bedside pleurodesis generally inappropriate. This outcome is surprising. Three months is a long time in the palliative care of cancer patients; survival differences in clinical trials are often measured in weeks. If breathing can be relieved usefully for even half of that time by a low-risk intervention, it is likely to be offered in practice. Defining a group of patients with less than three months to live was intended to focus attention on the relative merits of palliation, not to write the group off as being so near death as to preclude intervention for relief of breathlessness. Did the choice of three months life expectancy in the indication list send out a different message? Were experts interpreting the time scale differently? Three months was perhaps not a useful threshold.

Severity of breathlessness did not appreciably influence the appropriateness rating for bedside pleurodesis. For surgical pleurodesis there was a strong negative trend: the worse the breathlessness the less appropriate the intervention was deemed. This is incongruous. The primary purpose of pleurodesis is to relieve breathlessness: the worse the breathlessness the bigger the beneficial effect. The degree of breathlessness was intended to be a signifier of the ability of a patient to benefit from the procedure. Non surgical panellists used the dyspnoea score to determine fitness for surgery.

A strength of the study was that we had a chair very experienced in the method and who followed it to the letter, so we believe that we gave it a fair trial. Nevertheless, we identified two major conclusions of the panel that were not in keeping with the evidence from the systematic review with which they were provided and out of kilter with clinical sense. That our panellists were first timers is an inherent weakness of RAM, for that is the way most panels are composed.

LESSONS LEARNED

RAM aims to convert categorized patient characteristics and panellists' quantified judgments into a practical tool to aid decision making.[11,12] The method is complex and time consuming. The number of combinations and permutations made the process unwieldy to the point of being overwhelming. Most expert panel reports produce hundreds of scenarios and their associated ratings. During the panel meeting it became clear that clinicians were so used to picturing whole patients in their minds' eye that they found it difficult (for some impossible) to deconstruct the factors that would lead them to one decision rather than another.

The method requires experts to use consistently values that dictate the relative weight attached to each clinical dimension. There is a limit to the number of facts about a hypothetical patient that can be juggled in the mind. Beyond about seven dimensions, panellists are likely to pay attention to only one or two and begin to use inconsistent judgements, or to take 'short-cuts' that reduce the cognitive workload.[13]

It would have been possible for clinical parameters to be much more tightly defined and their intended purpose spelled out but at some point the exercise would become redundant because we would lead the clinicians towards what we consider to be the most desirable result. For example, it was a varying interpretation of the intended significance of breathlessness that led to confusion. If to avoid this we had put 'breathlessness meriting relief' we force the decision one way. If we put 'not fit for surgery' we close a gate on that pathway and force the decision another way; it shuts out all other considerations.

Absence of tissue diagnosis tended to 'trump' other considerations. Histological proof of cancer is generally regarded as important for all subsequent management decisions. When there was no tissue diagnosis, surgical pleurodesis gives an ideal opportunity to obtain biopsies, while bedside pleurodesis may be excluded if the diagnosis is unproven.

We are not alone in finding the process flawed. In a detailed study of the RAM, Raine *et al.* constructed 16 parallel panels in a complex prospective controlled design. They concluded 'A formal consensus development method produced judgements that were consistent with our assessments of the research evidence in about half the scenarios considered.' So half the statements were not in accord with evidence.[14] In spite of nearly two decades of work promulgating this approach[8,15] we do not believe it has been sufficiently critically appraised.

Individual clinicians are highly influenced by memorable adverse events and will change their practice contrary to evidence.[16] As the judgments of expert panels come under ever greater scrutiny, and as panels consider clinical areas where trials are lacking or absent, it is increasingly found that expert panellists make decisions incongruous with existing research evidence or clinical practice.[17] Revisiting these problem areas, at an interval, should be an inherent part of the process. Laboratory scientists using a new bioassay may not get it to work first time—why should expert panellists meeting just once expect the measures they generate to be any different?

Who should be the experts on the panel? All tasks are performed best by those with the right aptitudes. We chose eight clinicians largely on the basis of clinical experience, but the knowledge and skills that a clinician employs to treat an individual patient are not the same as those involved in RAM, which requires panellists to deconstruct the decision making process. Expertise in a clinical field may be an insufficient qualification and perhaps evidence of ability to analyze decision-making should be a prerequisite for inclusion in a panel. There is an illusion that qualitative research, in the form of an expert panel processes, will just come naturally. Being an expert is not the same as being an expert panellist and maybe it is time to give thought to how potential panellists may be selected and trained.

And whose life is it anyway? How well is a panel of doctors equipped to know what patients want?[18–20] Currently, a flaw in doctor-based consensus methods is the lack of patient input. It is well recognized that the weight patients put on a symptoms varies widely. One may prefer to be left alone, preferring to tolerate their breathless, while another may grasp an opportunity to be just that bit more mobile and independent. At the very least this means that the appropriateness ratings can only inform, not determine, the decision making process.

The RAM offers an important attempt at articulating links between knowledge and judgement. There are pitfalls for the unwary: the 'trumping' effect of the need for a tissue diagnosis, the double play of breathlessness, and the judgement about prognosis, were all revealed in this experience. There are variations in opinion based on the same evidence, all face to face consensus processes can be hijacked by rhetoric, and there are wide gaps between the evidence and what doctors actually do. The method can—and should—evolve, with consideration given to selection and training of panellists and the need for panel iteration.[15] We need to understand and refine the method and improve it if we are obliged to play by its rules.

Acknowledgment We are grateful to colleagues Willie Fountain, Robert Cameron, Robert Davies, Robin Rudd, Nihal Shah, Alex West and Bernie Foran who worked on the panel.

REFERENCES

1 Guyatt G, Gutterman D, Baumann MH, *et al*. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians task force. *Chest* 2006;**129**:174–81

2 Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ* 2003;**327**:1459–61

3 Tan C, Treasure T, Browne J, Utley M, Davies CWH, Hemingway H. Appropriateness of VATS and bedside thoractostomy talc pleurodesis as judged by a panel using the Rand/UCLA appropriateness method (RAM). *Interactive Cardiovascular and Thoracic Surgery* 2006;**Mar**:doi 10.1510/icvts.2005.123919

4 Sahn SA. State of the art. The pleura. *Am Rev Respir Dis* 1988;**138**:184–234

5 Dresler CM, Olak J, Herndon JE, *et al*. Phase III intergroup study of talc poudrage vs talc slurry sclerosis for malignant pleural effusion. *Chest* 2005;**127**:909–15

6 Shaw P, Agarwal R. Pleurodesis for malignant pleural effusions. *Cochrane Database Syst Rev* 2004;**CD002916**

7 Tan C, Sedrakyan A, Browne J, Swift S, Treasure T. The evidence on the effectiveness of management for malignant pleural effusion: a systematic review. *Eur J Cardiothorac Surg* 2006;**29**:829–38

8 Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care* 1986;**2**:53–63

9 Hemingway H, Banerjee S, Timmis A. Using guidelines for coronary revascularisation: how many are needed and are they any good? *Heart* 2000;**83**:5–6

10 Hemingway H, Crook AM, Banerjee S, *et al*. Hypothetical ratings of coronary angiography appropriateness: are they associated with actual angiographic findings, mortality, and revascularisation rate? The ACRE study. *Heart* 2001;**85**:672–9

11 Wietlisbach V, Vader JP, Porchet F, Costanza MC, Burnand B. Statistical approaches in the development of clinical practice guidelines from expert panels: the case of laminectomy in sciatica patients. *Med Care* 1999;**37**:785–97

12 Stoevelaar HJ, McDonnell J, Stals H, Smets L. Gastro-protective treatment in patients using NSAIDs. Development of appropriateness criteria by a multidisciplinary expert panel. *Scand J Rheumatol* 2003;**32**:162–7

13 Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 1956; **63**:81–97

14 Raine R. An experimental study of the determinants of group judgements in clinical guideline development. *Lancet* 2004;**364**:429–37

15 Raine R, Sanderson C, Black N. Developing clinical guidelines: a challenge to current methods. *BMJ* 2005;**331**:631–3

16 Choudhry NK, Anderson GM, Laupacis A, Ross-Degnan D, Normand SL, Soumerai SB. Impact of adverse events on prescribing warfarin in patients with atrial fibrillation: matched pair analysis. *BMJ* 2006; **332**:141–5

17 Glasier A, Brechin S, Raine R, Penney G. A consensus process to adapt the World Health Organization selected practice recommendations for UK use. *Contraception* 2003;**68**:327–33

18 Dowie J, Wildman M. Choosing the surgical mortality threshold for high risk patients with stage Ia non-small cell lung cancer: insights from decision analysis. *Thorax* 2002;**57**:7–10

19 Treasure T. Whose lung is it anyway? *Thorax* 2002;**57**:3–4

20 Hamel MB, Goldman L, Teno J, *et al*. Identification of comatose patients at high risk for death or severe disability. SUPPORT Investigators: Understand Prognoses and Preferences for Outcomes and Risks of Treatments. *JAMA* 1995;**273**:1842–8