

Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive

The large and diverse nature of guideline committees can make consensus difficult.

Roman Jaeschke and colleagues describe a simple technique for clarifying opinion

Guidelines have become an important vehicle for influencing clinical practice. Many local, national, and international societies now go through the process of identifying relevant clinical areas, formulating specific clinical questions, reviewing the applicable evidence, and formulating recommendations that they believe clinicians and their patients should follow.

Over the years, in recognition of the diversity of individuals required to produce optimal recommendations (content experts, methodologists, front line clinicians, patients' representatives), guideline panels have grown in size. The resulting large and diverse panels present challenges for decision making, such as ensuring that all participants have a voice and can influence the results of the debate, ensuring transparency, dealing with disagreement, achieving consensus, and resolving situations in which consensus is not possible.

Guideline panels often use only informal processes to deal with these challenges. Informal processes are, however, vulnerable

to the idiosyncrasies of small or moderate sized group interaction. Factors including time pressure; fatigue; lack of expertise in content, methods, or group leadership; and, most importantly, dominance by individuals with powerful personalities and intimidating reputations threaten the integrity of the process.

Those interested in the science of guideline development have developed two strategies to deal with these problems. The first uses structured approaches to collect, analyse, and summarise the relevant evidence and to use that evidence to produce and grade recommendations. These approaches are epitomised by the method suggested by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, which has developed an increasingly widely adopted structure for developing guidelines.¹⁻⁶ The second relies on somewhat formalised processes to encourage a consensus to which all panellists can contribute more or less equally.⁷⁻⁸

In this article, we briefly review consensus development techniques,⁹ describe a quality improvement and guideline development group (the Surviving Sepsis Campaign), and introduce the GRADE grid—an instrument recently developed and implemented by the Surviving Sepsis Campaign for use within the GRADE approach.

Formal consensus development techniques

The most popular techniques for developing consensus are the Delphi method, the nominal group technique, and a combination of these two approaches. The Delphi method, which was originally used to forecast the influence of technology on warfare, systematically gathers opinion from a number of stakeholders or experts. Large numbers of participants can be included in this process, during which contributors answer questionnaires in two or more rounds, usually working independently without meeting in person. After each round, a facilitator provides an anonymous summary of the contributors' opinions from the previous



GRADE grid for recording panellists' views in development of guidelines (including examples of propositions from the Surviving Sepsis Campaign and number of panellists who voted for each option)

	GRADE score				
	1	2	0	2	1
Balance between desirable and undesirable consequences of intervention	Desirable clearly outweigh undesirable	Desirable probably outweigh undesirable	Trade-offs equally balanced or uncertain	Undesirable probably outweigh desirable	Undesirable clearly outweigh desirable
Recommendation	Strong: "definitely do it"	Weak: "probably do it"	No specific recommendation	Weak: "probably don't do it"	Strong: "definitely don't do it"

For each proposition below, please mark with an "X" the cell that best corresponds to your assessment of the available evidence, in terms of benefits versus disadvantages

Use of (as opposed to no use of):	1	2	0	2	1
Low dose steroids in patients with septic shock responsive to fluids and vasopressors	0	5	4	8	4
Low dose steroids in patients with septic shock poorly responsive to fluids and vasopressors	5	16	0	0	0
SDD in ventilated patient (local and systemic)	0	9	4	8	1
rhAPC in patients with septic shock and high risk of death	6	15	1	0	0

SDD=selective digestive decontamination, rhAPC= recombinant human activated protein C.

*Participants were provided with guidance on factors to be taken into account in formulating a recommendation (box 1) and the implications of strong versus weak recommendations (box 2).

round, as well as the reasons they provided for their judgments. Participants are encouraged to revise their earlier answers in light of the replies from other members of the group. In general, during this process the range of the answers decreases, and the group converges towards a common answer. The process terminates after a predefined stop criterion (such as number of rounds, achievement of consensus, stability of results).^{9 10}

The nominal group technique elicits opinions from a smaller number of experts who meet in person. Each person is given equal opportunity to speak, and there is formal feedback by the organisers to the participants, structured face to face interactions, periods of private (non-interacting) activity such as development of ideas or ranking opinion, and an explicit method for final resolution. One method of resolution involves definition of several options that are ranked from most to least acceptable by all participants.

Both these techniques are used in a variety of situations where consensus needs to be built and not just for guidelines. For example, they have been shown to be valuable in establishing national research priorities¹¹ and in developing international training programmes.¹² Modifications of these methods are common—for example, voting on options in the nominal group technique rather than ranking—and

each technique can vary in design and implementation. Other methods, specific for guidelines developers, have been proposed.^{9 13}

Surviving Sepsis Campaign

Over 50 experts from more than 10 countries participated in the development of guidelines on managing severe sepsis and septic shock as part of the Surviving Sepsis Campaign.¹⁴ The first edition of the campaign's guidelines was published in 2004 and the most recent in 2008. The 2008 guidelines were developed using the GRADE approach to classify the quality of underlying evidence and the strength of recommendations.¹ GRADE classifies quality of evidence as high, moderate, low, or very low. The system allows the quality of evidence derived from observational data to be upgraded from low to moderate or high categories and the quality of evidence coming from randomised trials to be downgraded depending on the details of design and execution of such studies. The approach to deciding on the quality of evidence, while in its optimal application highly structured, nevertheless requires subjective judgment and thus invites differences of opinions.

Subjective judgment is also involved in classifying recommendations as strong or weak. The guideline panel has to decide whether the desirable effects of adherence

to a recommendation will outweigh the undesirable effects, and the strength of a recommendation reflects the group's degree of confidence in that assessment. A strong recommendation in favour of an intervention reflects the collective judgment that the desirable effects of the intervention (beneficial health outcomes, less burden on staff and patients, and cost savings) will clearly outweigh the undesirable effects (harms, more burden, and greater costs). A weak recommendation reflects the collective opinion that the desirable effects will outweigh the undesirable effects but the panel

Box 1 Factors that influence the strength of recommendation

Balance between desirable and undesirable effects—The larger the difference between the desirable and undesirable effects, the more likely a strong recommendation is warranted. The narrower the gradient, the more likely a weak recommendation is warranted

Quality of evidence—The higher the quality of evidence, the more likely a strong recommendation is warranted

Values and preferences—The more variability in values and preferences, or more uncertainty in values and preferences, the more likely a weak recommendation is warranted

Costs (resource allocation)—The higher the costs of an intervention (that is, the more resources consumed) the less likely a strong recommendation is warranted

is not confident about the trade-offs—either because key evidence is of low quality (and thus the benefits and risks remain uncertain) or because the benefits and downsides are closely balanced.

The Surviving Sepsis Campaign recognised the need for a more formal consensus process for resolving disagreement, interpreting evidence, and making recommendations, particularly in a climate of rapid change with new information emerging from ongoing clinical trials. This need was highlighted by criticism of the previous iteration of the campaign's guidelines.¹⁵ This criticism focused on conflict of interest and alleged manipulation of the academic authors by the drug industry.

Campaign consensus development

The consensus development techniques used by the campaign members and executive committee included a plenary consensus conference (the original meeting of all participants and organisations); small specialist working groups on each specific issue or intervention; two sequential modified nominal group meetings involving 15-30 people that considered the output from the working groups; and a modified Delphi method involving the whole group in iterative discussion by email. The primary area of disagreement that emerged during these processes was the strength of particular recommendations. Difficulties achieving consensus highlighted the need for a more formal approach to resolving disputes.

The campaign therefore decided on a voting procedure for the nominal group meetings guided by the following rules. Firstly, in areas of continuing disagreement, a recommendation for or against a particular intervention (compared with a specific alternative) required at least 50% of participants in favour, with less than 20% preferring the comparator (the options could be judged equal). Failure to meet this criterion resulted in no recommendation. Secondly, for a recommendation to be graded as strong rather than weak, at least 70% of participants were required to endorse it as strong.

Application of GRADE grid

To explore the range and distribution of the opinions held by panel members within the GRADE framework, we designed and implemented the GRADE grid (table). The grid allows members of the consensus panel to record their views about the balance between the benefits and disadvantages (downsides) of specific interventions, based on their analysis of the available evidence. This assessment is then mapped to the strength of recommenda-

Box 2 Examples of implications of strong and weak recommendations

Strong recommendation for intervention

For patients—Most people in this situation would want the recommended course of action and only a small proportion would not

For clinicians—Most people should receive the intervention

For quality monitors—Adherence to this recommendation could be used as a quality criterion or performance indicator. If clinicians choose not to follow such a recommendation, they should document their rationale

Weak recommendation for intervention

For patients—Most people in this situation would want the suggested course of action, but many would not

For clinicians—Examine the evidence or a summary of the evidence yourself and be prepared to discuss that evidence with patients, as well as their values and preferences

For quality monitors—Clinicians' discussion or consideration of the pros and cons of the intervention, and their documentation of the discussion, could be used as a quality criterion.

No specific recommendation

The advantages and disadvantages are equivalent

The target population has not been identified

Insufficient evidence on which to formulate a recommendation

tion for the use, or not, of each intervention. Each proposition is expressed in a neutral manner ("Use of . . .").

To guide their use of the grid, all participants received instructions describing factors that influence the strength of recommendation and the implications of strong and weak recommendations (boxes 1 and 2). Each vote dealt with a clinical problem presented as a proposition with which panellists could express varying degrees of approval or disapproval. Panellists completed the form after full restatements of the proposition (explicit description of population, intervention, comparator, and outcomes), presentation of the evidence, and review of potential sources of disagreement (box 1) as perceived by the leaders of the debating parties.

Examples of the process

Participants, already well informed about the GRADE method, found the form easy to use. The introduction of the task, instructions, and answering related questions took less than 10 minutes. After agreement on the proposition and presentation of the relevant evidence, completing the form for each recommendation took less than two minutes. Support staff tabulated the votes and presented the results to the group. The following examples highlight how, in retrospect, the GRADE grid was helpful.

Clarifying decisions

In the case of steroid supplementation in septic shock, two propositions were explored to clarify the opinions of the participants. The first proposition dealt with use of steroids (versus not using them) in adult patients with septic shock resistant to

initial treatment with fluids and vasopressors (drugs raising blood pressure). The second proposition dealt with steroid use in adult patients with septic shock who responded to initial treatment. The original proposal from the steroid subcommittee was to provide a strong recommendation to use steroids in the first group (blood pressure unresponsive to fluid and vasopressors) and a strong recommendation not to use them in the second (blood pressure responsive to fluids and vasopressors). Members of the full committee challenged this proposal when it was presented to them electronically because of the difficulty of making two strong recommendations for and against use while being unable to define responsiveness to treatment precisely. We therefore used the poll-



Use of the grid by the Surviving Sepsis Campaign facilitated rapid achievement of consensus and closure on topics that had previously generated extended but apparently inconclusive discussion need

Demonstrating patterns of uncertainty

Selective digestive decontamination (use of prophylactic antibiotics to prevent infection in ventilated patients) remains controversial despite extensive research validation. It became evident in plenary discussion that consensus would not be obtained without a formal voting process. The table shows the degrees of uncertainty about the potential effect of this treatment, with participants polling equally for or against its use on a weak recommendation, and a substantial proportion undecided. Since 50% or more of the panel could not agree on a direction of recommendation, the committee therefore chose not to make a recommendation for or against its use. The result of the vote effectively closed further discussion, which might otherwise have continued for a long time.

Decisions about strength of recommendation

In case of activated protein C, the original meeting of the panel and subsequent email discussion concerning the choice of a strong versus weak recommendation had not led to a solution. This discussion was effectively put to rest by polling using the grid, which showed that the majority preferred a weak recommendation in favour of its use in patients with a clinical assessment of high risk of death (table). This result was accepted unanimously by the whole panel without requiring further discussion.

Conclusions

The most challenging part of this consensus process was the precise definition of acceptable clinical questions (propositions), including population, intervention, and comparator, and the need to structure the proposition in a neutral way that allowed the full range of options. In situations where consensus is elusive, once the guideline panel has formulated the precise clinical question or questions, we propose the use of a structured approach to explore views on balance between the desirable and undesirable consequences of an intervention. The GRADE grid described here provides a useful and efficient way to examine the range of opinions which inform further discussion and then permits polling within the group.

Use of the grid by the Surviving Sepsis Campaign facilitated rapid achievement of

consensus and closure on topics that had previously generated extended but apparently inconclusive discussion among expert participants with vigorous views on both the science and the interpretation of research evidence. The validity of our positive opinion may be limited by the fact that most of us participated in generating the campaign guidelines and the voting process.

Voting rules were specific to the campaign's work. We chose to maintain anonymity of voting, as this provides the best opportunity for free expression of views. Open voting could perhaps restrain voting behaviour driven by conflict of interest. However, we believe that private voting using the grid combined with careful constitution of the nominal group will ensure that such conflicts (where they exist) are balanced or their impact minimised.

Although preparing high quality GRADE evidence summaries requires extensive resources, use of the grid does not. Indeed, our impression is that the grid results in efficiencies through the rapid and explicit clarification of panellists' views, and the extent of agreement and disagreement. We believe that the grid may be helpful for any guideline group using the GRADE approach to achieve consensus or to understand the patterns of uncertainty that surround the interpretation of scientific evidence.

Roman Jaeschke clinical professor, Department of Medicine, McMaster University, Hamilton, ON, Canada L8P 3B6

Gordon H Guyatt professor, Department of Medicine, McMaster University, Hamilton, ON, Canada L8P 3B6
Department of Clinical Epidemiology and Biostatistics, McMaster University

Phil Dellinger professor, Division of Critical Care, Cooper University Hospital and Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Camden, NJ, USA

Holger Schünemann professor, Department of Epidemiology, Italian National Cancer Institute Rome, Rome 00144, Italy

Mitchell M Levy professor, Division of Pulmonary and Critical Care Medicine, Brown University School of Medicine, Providence, RI, USA

Regina Kunz associate professor, Basle Institute of Clinical Epidemiology, University Hospital Basle, 4031 Basle, Switzerland

Susan Norris assistant professor, Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239, USA

Julian Bion professor of intensive care medicine for the GRADE working group, University of Birmingham, Queen Elizabeth Hospital, Birmingham B15 2TH

Correspondence to: J Bion J.F.Bion@bham.ac.uk

Accepted: 19 May 2008

We thank the Surviving Sepsis Campaign guidelines

development group for the use of this material.

Contributors: The other members of the GRADE working group were Morio Aihara, Jeff Andrews, Jan Brožek, Jonathan Craig, Benjamin Djulbegovic, Signe Flottorp, Yngve Falck-Ytter, Suzanne Hill, Merce Marzo, Andy Oxman, Bob Phillips, Arturo Salazar, and John Williams. RJ, JB, GHG, and PD developed the concept of GRADE grid and used this instrument to develop practice guidelines. All authors participated in interpretation of observations and drafting and revising the manuscript. All authors approved the final version. RJ is guarantor.

Competing interests: GHG, HS, RK, and RJ receive honoraria and consulting fees for activities in which their work with GRADE is relevant. HS is documents editor and methodologist for the American Thoracic Society; one of his roles in these positions is helping implement the use of GRADE. He supports the implementation of GRADE by organisations worldwide. JB is a past member of the executive of the Surviving Sepsis Campaign. Occasional consulting fees or honoraria are donated to his department and are unrelated to either the Surviving Sepsis Campaign or GRADE.

Provenance and peer review: Not commissioned; externally peer reviewed.

- Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605-14.
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-6.
- Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Schünemann HJ. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008;336:995-8.
- Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Gunn EV, Liberati A, et al. Going from evidence to recommendations. *BMJ* 2008;336:1049-51.
- Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading the quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106-10.
- Guyatt GH, Oxman AD, Kunz R, Jaeschke R, Helfand M, Vist GE, et al. Incorporating considerations of resources use into grading recommendations. *BMJ* 2008;336:1170-3.
- Fretheim A, Schünemann HJ, Oxman AD. Improving the use of research evidence in guideline development: 3. Group composition and consultation process. *Health Res Policy System* 2006;4:15.
- Fretheim A, Schünemann HJ, Oxman AD. Improving the use of research evidence in guideline development: 5. Group processes. *Health Res Policy System* 2006;4:17.
- Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;2.
- Thangaratnam S, Redman CWE. The Delphi technique. *Obstetrician Gynaecologist* 2005;7:120-5.
- Vella K, Goldfrad C, Rowan KM, Bion JF, Black NA. Use of consensus development to establish national research priorities in critical care. *BMJ* 2000;320:976-80.
- CoBaTriCE Collaboration. Consensus development of an international competency-based training programme in intensive care medicine. *Intensive Care Med* 2006;32:1371-83.
- Raine R, Sanderson C, Black N. Developing clinical guidelines: a challenge to current methods. *BMJ* 2005;331:631-3.
- Dellinger RP, Levy MM, Carlet JM, Bion J, Parker MM, Jaeschke R, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Crit Care Med* 2008;36:296-327.
- Eichacker PQ, Natanson C, Danner RL. Surviving sepsis—practice guidelines, marketing campaigns, and Eli Lilly. *N Engl J Med* 2006;355:1640-2.

Cite this as: *BMJ* 2008;337:a744