Users' Guides to the Medical Literature

IX. A Method for Grading Health Care Recommendations

Gordon H. Guyatt, MD; David L. Sackett, MD; John C. Sinclair, MD; Robert Hayward, MD; Deborah J. Cook, MD; Richard J. Cook, PhD; for the Evidence-Based Medicine Working Group

THE ULTIMATE PURPOSE of applied health research is to improve health care. Summarizing the literature to adduce recommendations for clinical practice is an important part of the process. Recently, the health sciences community has reduced the bias and imprecision of traditional literature summaries and their associated recommendations through the development of rigorous criteria for both literature overviews[1-3] and practice guidelines.[4,5] Even when recommendations come from such rigorous approaches, however, it is important to differentiate between those based on weak vs strong evidence. Recommendations based on inadequate evidence often require reversal when sufficient data become available,[6] while timely implementation of recommendations based on strong evidence can save lives.[6] In this article, we suggest an approach to classifying strength of recommendations. We direct our discussion primarily at clinicians who make treatment recommendations that they hope their colleagues will follow. However, we believe that any clinician who attends to such recommendations would benefit from the increased understanding they will gain through reading this article.

GRADING HEALTH CARE RECOMMENDATIONS: PREVIOUS CRITERIA

In 1979, the Canadian Task Force on the Periodic Health Examination made one of the first efforts to specify the strength of practice recommendations.[7] This group classified the quality of the evidence regarding the benefit of interventions into one of four categories based on the quality of the individual study designs. Their classification of the strength of their recommendations was considerably less explicit, only labeling evidence as "good," "fair," or "poor." The original Canadian Task Force approach, with minor modifications, has been reaffirmed by the Canadian Task Force[8] and endorsed by the US Preventive Services Task Force.[9] Both task forces contributed to progress in developing ways of grading the strength of health care recommendations that enhance both their interpretability and validity.

ADVANCES IN METHODOLOGY

The classification system we present in this article is driven by four advances in translating evidence from original studies into clinical recommendations. First, methodologists have developed standardized approaches to the scientific conduct of literature reviews, and reviewers are increasingly using these approaches. This methodology includes systematic procedures and statistical techniques for combining results from different studies to minimize bias and increase precision.[10] Second, we have distinguished between clinical importance and statistical significance and realize that an intervention may be beneficial, but the effect too small to make the intervention worth administering.[11] The third advance is the more explicit acknowledgment that the strength of health care recommendations should depend on the precision of the estimated intervention effects: in general, the greater the sample size, the more precise our estimates of intervention effects, the narrower the confidence interval (CI) around our estimate of those effects, and the greater our ability to make strong recommendations. Finally, we are more aware that we may serve individual patients or groups of patients best if we withhold treatment for those at very low risk of clinical events while at the same time recommending treatment to those at higher risk.[12-15]

The Canadian and US Task Force criteria do not incorporate these advances. Members of our group have previously developed and modified criteria that addressed systematic overviews, but we failed either to clearly separate study design from the magnitude of the intervention effect, or to consider the impact of degree of patients' risk on treatment recommendations.[16,17] The approach we present in this article builds on the extensive work undertaken to date. We will focus on situations where investigations provide data regarding the effect of interventions on clinically important outcomes, whether the interventions are therapeutic, preventative, or diagnostic.

Our approach begins with the identification of a systematic overview of the existing evidence. By "systematic" we mean one that meets the following standards: the overview (1) addresses a focused clinical question; (2) uses appropriate criteria to select studies for inclusion; (3) conducts a comprehensive search; and (4) appraises the validity of the individual studies in a reproducible fashion. These standards are the same as those we recommend that clinicians use to identify an overview that is likely to yield an unbiased estimate of treatment effect.[18] Recommendations intended to influence clinical practice should be based on a current overview that meets these criteria.

COMPONENTS OF THE APPROACH TO GRADES OF RECOMMENDATION

In our framework, making a recommendation about a health care interven-

From the Departments of Medicine (Drs Guyatt, Hayward, and D. J. Cook) and Clinical Epidemiology and Biostatistics (Drs Guyatt, Hayward, D. J. Cook, and Sinclair), and the Department of Pediatrics (Dr Sinclair), McMaster University, McMaster University Faculty of Health Sciences, Hamilton, Ontario; the Centre for Evidence-Based Medicine, Nuffield Department of Medicine, University of Oxford (England) (Dr Sackett); and the Department of Statistics and Actuarial Sciences, Faculty of Mathematics, University of Waterloo (Ontario) (Dr R. J. Cook).

A complete list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Eric Bass, MD, MPH; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Andreas Laupacis, MD, MSc; Virginia Moyer, MD, MPH; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

Table 1.—Grades of Recommendations for a Specified Level of Baseline Risk*

A1	RCTs, no heterogeneity, CIs all on one side of threshold NNT
A2	RCTs, no heterogeneity, CIs overlap threshold NNT
B1	RCTs, heterogeneity, CIs all on one side of threshold NNT
B2	RCTs, heterogeneity, CIs overlap threshold NNT
C1	Observational studies, CIs all on one side of threshold NNT
C2	Observational studies, CIs overlap threshold NNT

*RCT indicates randomized controlled trial; CI, confidence interval; and NNT, number needed to treat to avoid one unwanted outcome.

Table 2.—Number Needed to Treat

	Bleeding Risk if Untreated, U	Relative Risk Reduction, (U−T)/U	Bleeding Risk if Treated, T	Absolute Risk Reduction, U−T	No. Needed to Treat, 1/(U−T), to Prevent a Bleed
Critically ill patient receiving mechanical ventilation and/or has a coagulopathy	0.037	58%	0.0155	0.0215	45
Critically ill patient breathing spontaneously without a coagulopathy	0.0014	58%	0.0006	0.0008	1250

tion requires the integration of three elements: the strength of the evidence presented in the overview; the threshold or magnitude of intervention effect at which benefit exceeds the risks of therapy, including both adverse effects and costs; and the relationships between the estimate of the magnitude of the intervention effect, the precision of that estimate, and the threshold. We will deal with each of these components in turn. In describing results of studies, we will consider the effect of the intervention on the clinical event that it is designed to prevent, which we will call the "target event." We will focus on the following: (1) the relative risk (RR), which is the ratio of the risk of target events in treated patients to the risk of target events in the untreated patients, and the RR reduction, or (1 − RR)[19]; (2) the absolute risk reduction, which is the difference in the absolute risk of the target event between treatment and control groups; and (3) the number needed to treat (NNT), which is the number of patients one needs to treat to prevent one target event (arithmetically, the inverse of the absolute risk reduction).[20]

Component 1: The Strength of the Evidence

Randomized Controlled Trials.—Because no other study design can provide the safeguards against bias associated with randomization, randomized controlled trials (RCTs) yield stronger evidence than other study designs. Overviews of RCTs, therefore, provide far stronger evidence than do overviews of cohort and case-control studies. The strength of evidence from an otherwise systematic overview of RCTs will, however, depend on the consistency of the results from study to study. When different studies in the same overview yield very different estimates of treatment effect (a situation we refer to as "heterogeneity" of study results), one must question why. Possibilities include differences in patient populations, the way the interventions were administered, the way the outcomes were measured, the way the studies were conducted, or the play of chance.[21,22] A statistical test of the homogeneity of the intervention effect asks the question, "Are the differences in treatment effect from study to study greater than one would expect simply as a result of chance"?

If investigators conducting an overview conclude that treatment has a different effect depending on the population or the way the intervention is administered, they may conduct separate overviews for the different populations or treatments.[21,22] When differences in treatment effect across studies are greater than one would expect by chance alone, and varying populations, interventions, outcomes, or study methods cannot explain the differences, inferences become weaker. We therefore rank the strength of evidence from overviews of RCTs according to the presence or absence of unexplained differences in results from study to study (Table 1). We rank overviews with significant and important heterogeneity (level B) lower than those without significant and important heterogeneity (level A).

Before concluding that recommendations be classified as level B rather than level A, we should be confident that the degree of heterogeneity is clinically important. Heterogeneity can be considered clinically important if there is a large difference in RR reduction across studies. If the estimates from the individual studies are imprecise, however, an apparent large difference may be due to the play of chance. We propose the following criteria for clinically important heterogeneity:

1. The difference in the estimate of RR reduction between the two most disparate studies is greater than 20% (for instance an RR reduction of 40% in one study and less than 20% in another).
2. The difference between the boundaries of the CIs between the two most disparate studies is greater than 5% (for instance, the lower boundary [the smallest RR reduction compatible with the data] in the first study is 30% and the upper boundary of the CI [the largest RR reduction compatible with the data] in the second study is less than 25%).

Before heterogeneity bears on the strength of treatment recommendations, it must be both clinically important and statistically significant ($P<.05$).

Observational Studies.—Because the potential for bias is much greater in cohort and case-control studies than in RCTs, recommendations from overviews combining observational studies will be much weaker.[23,24] Thus, we classify observational studies as providing weaker evidence than RCTs (Table 1).

Component 2: How Big an Impact of Treatment Warrants Its Use?

Any decision about initiating a preventive or therapeutic regimen represents a trade-off between patient or public benefits, on the one hand, and toxicity, cost, and administrative burden to patients and providers on the other. Clinicians do not, therefore, administer all effective treatments (effective in that they have a positive effect on some important outcome) to all potentially eligible patients. For example, H_2 receptor antagonists reduce the RR of serious bleeding in critically ill patients by approximately 58%.[25] However, a patient who is breathing spontaneously without a coagulopathy has a risk of serious bleeding of only 0.14% without treatment.[26] This baseline risk is so low that most clinicians would not consider it worth treating to lower the RR by another 58% (to 0.06%).

For administration of H_2 receptor antagonists to critically ill patients, and indeed for any treatment of any condition, it is useful to think of a threshold effect, above which one would treat and below which one would not. Moreover, it is informative to think of the number of patients one would need to treat to prevent a single serious gastrointestinal bleed.[27,28] Consider a group of critically ill patients who are receiving mechanical ventilation or who have a coagulopathy and whose risk of bleeding is therefore increased to 3.7%.[25,26] Treating such patients with H_2 receptor antagonists, one reduces their RR by 58%, to 1.55%. In absolute terms, their risk has fallen 2.15% (Table 2). The reciprocal of this absolute risk reduction is the NNT. In this case, 45 patients must receive prophylaxis to prevent an episode of serious bleeding.

Table 3.—How to Calculate the Threshold Number Needed to Treat

This table outlines how we calculate the threshold number needed to treat (NNT), a complete description of which will appear in an article we are preparing for publication. In describing how to calculate a threshold NNT, we will use the following notation:

T-NNT: the threshold number needed to treat
$\text{Cost}_{\text{treatment}}$: the cost of treating one patient
$\text{Cost}_{\text{target}}$: the cost of treating one target event
Cost_{AE}: the cost of treating one adverse event, with a further subscript 1 or 2 denoting the first and second adverse effects
Rate_{AE}: the proportion of treated patients who suffer an adverse event (again, subscripts 1 and 2 denoting the two adverse events)
$\text{Value}_{\text{target}}$: the dollar value we assign to preventing one target event
Value_{AE}: the dollar value we assign to preventing one adverse event (again, subscripts 1 and 2 denoting the two adverse events)

The general approach for generating the threshold NNT is based on the concept that at this threshold the value of treatment inputs equals the value of treatment outputs; that is, the net cost of treating the number of patients one needs to treat to prevent one patient having the target event equals the net value of the adverse events prevented or caused by treating that number of patients. The value of the treatment inputs includes the following:

the cost of treating the number of patients that will comprise the threshold NNT: $(\text{Cost}_{\text{treatment}})(\text{T-NNT})$
plus
the cost of treating the adverse events attributable to treatment in the number of patients that will comprise the threshold NNT: $(\text{Cost}_{\text{AE}})(\text{Rate}_{\text{AE}})(\text{T-NNT})$
minus
the cost of treating one target event: $\text{Cost}_{\text{target}}$

The value of the outputs includes the following:

the dollar value assigned to the one target event prevented: $\text{Value}_{\text{target}}$
minus
the dollar value assigned to adverse events attributable to treatment: $(\text{Value}_{\text{AE}})(\text{Rate}_{\text{AE}})(\text{T-NNT})$

Thus, we have:

$$[(\text{Cost}_{\text{treatment}})(\text{T-NNT})] + [(\text{Cost}_{\text{AE}})(\text{Rate}_{\text{AE}})(\text{T-NNT})] - \text{Cost}_{\text{target}} = \text{Value}_{\text{target}} - [(\text{Value}_{\text{AE}})(\text{Rate}_{\text{AE}})(\text{T-NNT})]$$

Rearranging:

$$\text{T-NNT}\ [\text{Cost}_{\text{treatment}} + (\text{Cost}_{\text{AE}})(\text{Rate}_{\text{AE}})] - \text{Cost}_{\text{target}} = \text{Value}_{\text{target}} - (\text{T-NNT})[(\text{Value}_{\text{AE}})(\text{Rate}_{\text{AE}})]$$

And solving for threshold NNT:

$$\text{T-NNT} = (\text{Cost}_{\text{target}} + \text{Value}_{\text{target}})/[\text{Cost}_{\text{treatment}} + (\text{Cost}_{\text{AE}})(\text{Rate}_{\text{AE}})] + [(\text{Value}_{\text{AE}})(\text{Rate}_{\text{AE}})]$$

In the example we have used in the body of the article concerning the prevention of gastrointestinal bleeding, there are two adverse effects attributable to treatment that we must consider. The equation therefore becomes the following:

$$\text{T-NNT} = (\text{Cost}_{\text{target}} + \text{Value}_{\text{target}})/[\text{Cost}_{\text{treatment}} + (\text{Cost}_{\text{AE1}})(\text{Rate}_{\text{AE1}}) + (\text{Cost}_{\text{AE2}})(\text{Rate}_{\text{AE2}})] + [(\text{Value}_{\text{AE1}}) + (\text{Rate}_{\text{AE1}})\ (\text{Value}_{\text{AE2}})(\text{Rate}_{\text{AE2}})]$$

Substituting the figures from the body of the article yields the following:

$$\text{T-NNT} = (12\,000+3000)/[65+(10\,000)(0.0006)+(500)(0.015)]+[(3000)(0.0006)+(300)(0.015)]$$
$$\text{T-NNT} = 15\,000/[65 + 6 + 7.5] + [18 + 4.5]$$
$$\text{T-NNT} = 15\,000/101$$
$$\text{T-NNT} = 148.5$$

We believe that it is important to consider costs in deciding on the threshold NNT. Some clinicians may be uncomfortable with including costs. A model for calculating the threshold NNT that neglects costs would use the following formula:

$$\text{T-NNT} = 1/[(\text{Value}_{\text{AE1}})(\text{Rate}_{\text{AE1}}) + (\text{Value}_{\text{AE2}})(\text{Rate}_{\text{AE2}})]$$

In this equation the value of the adverse events is not the dollar value as in the model that includes costs, but the value of the adverse event in terms of the target event. That is, if we decided that the negative consequences of an adverse event was only one-tenth as great as the negative consequences of the target event, the value of that adverse event would be 0.1.

Consider again the first group of critically ill patients we've mentioned, those who are breathing spontaneously and who don't have a coagulopathy. Their risk of bleeding without treatment, which we call the "baseline" risk, is 0.14%, their risk with treatment is 0.06%, and one must treat 1250 such patients to prevent a serious bleed (Table 2).

Should we treat either, or both, of these patients? This decision involves generating a threshold NNT. If the patients' risk without treatment is high enough, and the NNT is below the threshold, we administer treatment. If the patient's risk without treatment is low enough, and the NNT is therefore above the threshold, we would not treat.

Generating the threshold NNT involves three steps. In the first step, we identify two sorts of undesirable events. One is the target event, and the other is the adverse effects attributable to treatment. To generate the threshold NNT, we must specify the costs we incur when we treat patients, the costs we save when we prevent the occurrence of the target event, costs that we might incur as a result of preventing the target event, and the costs we incur when we look after patients who suffer adverse events associated with treatment.

In considering the decision whether to administer prophylaxis for gastrointestinal bleeding, some of the costs we specify below are based on a detailed economic analysis from a hospital's point of view (D. Heyland, A. Gafni, D. Cook, G. H. Guyatt, unpublished data, 1995), while others are much more approximate estimates. In this case, the cost of administering ranitidine during a patient's 10-day stay in the intensive care unit (calculated, as are all our costs, based on Canadian data) is approximately \$65 (including drug costs and costs of administering the treatment) and the cost of treating a gastrointestinal bleed is \$12 000. Adverse effects of the H_2 receptor antagonist ranitidine include hepatitis with hepatic failure (an incidence of 0.06%,[29] with a treatment cost of \$10 000 per episode) and central nervous system toxicity (an incidence of 1.5%,[30] with a cost of \$500 per episode).

The second step in generating the threshold NNT is assigning relative values to the outcomes and relating them to dollar costs. These values may come from health workers, administrators, patients, or a large random sample of the general public and might use one of a number of approaches (such as individual interviews or a group consensus process) to assess utility.[31] While there is no consensus about either who should be deciding values or the best method of establishing that group's values, we would recommend individual interviews with either patients or the general public. Whatever population and approach to eliciting values one chooses, the process would involve (in this case) determining the degree of satisfaction, distress, or desirability that people associate with having an episode of gastrointestinal bleeding relative to an episode of liver toxicity or central nervous system toxicity. The process then involves deciding how much money should be allocated to prevent a single episode of gastrointestinal bleeding, which in turn sets the money we would be willing to spend to avoid the adverse events attributable to treatment.[32]

For purposes of the present discussion, we have not actually obtained values from a random sample of the population, but have guessed at what the population might say. In this case, we would be willing to spend \$3000 to prevent one gastrointestinal bleed. We have equated one episode of liver toxicity and 10 episodes of central nervous system toxicity to a serious gastrointestinal bleed. Thus, we would be willing to spend \$3000 to avoid one episode of liver toxicity and \$300 to avoid one episode of central nervous system toxicity. We explain the algebra involved in calculating the threshold NNT in Table 3; as it turns out, the figures above generate a threshold NNT of approximately 150.

Figure 1 presents the relationship between the treatment NNT, the thresh-

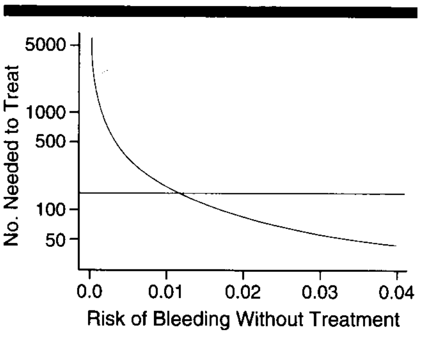

Figure 1.—Relationship between number needed to treat (NNT) associated with treatment, threshold NNT (horizontal line), and risk of bleeding without treatment for critically ill patients.

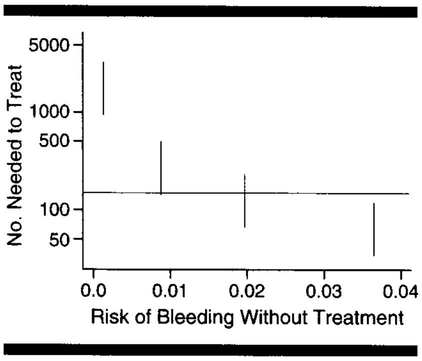

Figure 2.—Levels of baseline risk and threshold number needed to treat (NNT). Vertical lines represent the 95% confidence intervals around the treatment NNT at baseline risks of 0.14%, 0.9%, 2%, and 3.7%.

old NNT, and the risk of bleeding without treatment for critically ill patients. In constructing Figure 1, we have used the RR reduction we can expect with administration of H_2 receptor antagonists (58%), and the threshold NNT of 150 that we have generated. The horizontal line at an NNT of 150 represents this threshold NNT. The decreasing curve represents the NNT for any given risk of bleeding without treatment, which we will call the "treatment NNT line." Points on this line include the groups of patients from Table 2: patients with a risk of serious bleeding without treatment of 3.7%, for whom the NNT is 45, and patients with a risk of serious bleeding without treatment of 0.14%, for whom the NNT is 1250. The treatment NNT line crosses the threshold NNT at a risk without treatment of 1.15%. Therefore, our judgment is that treatment is warranted in patients whose risk of serious bleeding without treatment is greater than 1.15%, and not warranted for those whose risk is less than 1.15%.

The threshold NNT will vary depending on the values the clinician and patient place on its components. Some clinicians may be uncomfortable including costs as a consideration in the decision to treat. The strength of the threshold approach is that those recommending policy can, in generating a threshold NNT, make explicit the values they place on avoiding clinical events, adverse effects, and costs incurred or avoided, or omit costs from the consideration. In Table 3, we provide a method of calculating the threshold NNT without considering costs. Clinicians can examine the basis for the decision regarding threshold NNT, and the implications of differences in values, and the lower or higher threshold generated as a result of different values.

Component 3: How Much Does the Treatment Work?

A meta-analysis is a quantitative overview that yields the best estimate of the treatment effect by pooling results from different trials. This estimate is called a "point estimate" to remind us that although the true value lies somewhere in its neighborhood, it is unlikely to be exactly correct. Confidence intervals tell us the range within which the true treatment effect likely lies.[33,34] We usually (though arbitrarily) use the 95% CI, which can be interpreted as defining the range that would include the true treatment effect 95% of the time on repetition of the experiment.

Given a specified risk of a clinical event without treatment, we can use the reduction in RR of clinical events with treatment and the CI around that reduction in RR, to calculate not only the NNT, but also the CI around the NNT. The relationship between that CI and the threshold NNT will have a profound effect on the strength of any recommendation to treat or not to treat. There are four possible relationships between the threshold NNT, the point estimate of the treatment effect, and the CI around the point estimate. We will examine each of these four in turn.

Consider critically ill patients who are receiving mechanical ventilation or have a coagulopathy. We have already decided that since their NNT lies below the threshold, they should be treated with H_2 receptor antagonists (or some equivalent treatment) (Table 2, Figure 1). We must remember, however, the upper boundary of the CI around the NNT. This boundary represents the smallest reduction in risk and thus the largest NNT, which is likely to be consistent with the data. In this case, the 95% CI around the RR reduction of 58% ranges from 79% to 21%, and the corresponding CI around the NNT, given the risk without treatment of 3.7%, ranges from 34 to 129. Here, the boundary of the CI that represents the highest NNT consistent with the data is still less than the threshold NNT of 150. We can be confident that the treatment for patients whose risk of bleeding is 3.7% does more good than harm, on average, given the relative values and costs we have specified.

Consider critically ill patients who are neither receiving mechanical ventilation nor have a coagulopathy and whose risk of bleeding is therefore 0.14%. Given the 58% RR reduction, we must treat 1250 such patients to prevent a bleed (Table 2). The 95% CI around this NNT ranges from 904 to 3401. The boundary of the CI that represents the largest plausible treatment effect, and thus the smallest NNT (904), is greater than the threshold NNT of 150. We can therefore be confident that the risks and costs of treatment outweigh the benefits.

If the risk of bleeding without treatment is intermediate, the recommendation is less clear. Take, for instance, a critically ill patient with a bleeding risk of 2%. Given an RR reduction of 58%, we must treat 86 such patients to prevent a bleed. Given the range of the 95% CI around the RR reduction (79% to 21%), the true NNT may lie between 63 and 238. The boundary of the 95% CI that represents the smallest plausible treatment effect, and thus the greatest NNT, 238, is greater than the threshold NNT. While the overall recommendation will still be to treat patients with this level of risk of bleeding, our strength of inferences will be weaker.

Similarly, if one considers a patient with a risk of serious bleeding without treatment of 0.9%, the most likely NNT is 192, but the 95% CI ranges from 141 to 529. Since the most likely NNT is above the threshold, the recommendation will be to withhold treatment, but because the 95% CI overlaps the threshold NNT of 150, the strength of inference is relatively weak.

We present results from all four levels of baseline risk (0.14%, 0.9%, 2%, and 3.7%) together with the threshold NNT in Figure 2.

THE FINAL PRODUCT: RECOMMENDATIONS

If one combines the strength and heterogeneity of the primary studies with the magnitude and precision of the treatment effect as it relates to the threshold NNT, one can decide on the strength of the recommendation to treat or not to treat (Table 1). As we have demonstrated, the recommendation may change from "offer the intervention" when the baseline risk is high, to "don't offer the intervention" when the baseline risk is low. We believe that within RCTs, whether the CI on the NNT overlaps the threshold NNT is more important than the presence of heterogeneity. However, RCT evidence is always stronger than evidence from observational studies. Thus, for any given baseline risk, A1 and B1 designate the strongest recommendations, A2 and B2 represent intermediate-strength recommendations, and C1 and C2 are the weakest recommendations.

COMMENT

There are many issues in arriving at recommendations that remain to be fully explored. The .05 threshold for deciding whether or not heterogeneity is statistically significant, the proposed criteria for deciding whether heterogeneity is clinically important, and the choice of 95% for the CI around the treatment NNT are all arbitrary. Our choice of the 95% CI is based on tradition. Less stringent values would lead to narrower CIs (and thus more level 1 recommendations) and may ultimately be judged more appropriate.

The decision regarding the threshold NNT requires data both on costs and on the relative values we place on varying outcomes, data that will often not be available. Limitations in the data will emphasize the need to conduct additional rigorous studies. In the meanwhile, we must make treatment decisions and these decisions imply estimates of costs and values. Making these estimates explicit is worthwhile, even if we acknowledge their imprecision. We can examine the treatment implications of varying assumptions about costs and values (and thus, varying threshold NNTs). This emphasizes the absolute requirement to be explicit about what drives our decisions, particularly the underlying values.

The decision about the threshold NNT may vary in different practice settings and from patient to patient. We suggest that those making recommendations for clinical practice be explicit about how they arrive at their threshold NNT. They must consider all major toxicity, annoyance or inconvenience for the patient, the administrative burden on the health care system, and the cost of treatment, and describe how they have valued each component. If clinicians disagree with the values underlying a particular threshold NNT or work in a setting in which a particular threshold NNT does not apply, they can generate a new threshold NNT consistent with their values or practice setting. They could still use the overview evidence and the treatment NNT and quickly generate recommendations.

Our approach represents one in a series of steps along the road to optimal categorization of treatment recommendations and will likely require modification. Nevertheless, four elements of the approach presented here should help us move forward in the search for better ways of framing treatment recommendations. First, recommendations must be based on systematic overviews of methodologically sound primary studies. Second, those making recommendations must specify a threshold level of impact that warrants recommendation for applying the intervention. Third, recommendations will almost certainly vary when the magnitude of risk without treatment varies. Finally, recommendations must be based on two clearly separated components, the design and heterogeneity of the primary studies, on the one hand, and the magnitude and precision of the estimates of the treatment effects on the other. We hope that clinicians and policymakers find these insights useful in future development of treatment recommendations.

We would like to thank John Simes for his insightful comments, which helped us address a number of important conceptual issues. Kjell Apslund, Iain Chalmers, Peter Gotzsche, Chris Silagy, and Salim Yusuf all provided helpful comments on earlier drafts of the manuscript. We offer special thanks to Deborah Maddock who has provided outstanding administrative support and coordination for the activities of the Evidence-Based Medicine Working Group.

References

1. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med.* 1987;316:450-455.
2. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272:1367-1371.
3. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med.* 1987;107:224-233.
4. Hayward RSA, Wilson MC, Tunis SR, Bass EB, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA.* 1995;274:570-574.
5. Wilson MC, Hayward RSA, Tunis SR, Bass EB, Guyatt G, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help you in caring for your patients? *JAMA.* 1995;274:1630-1632.
6. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA.* 1992;268:240-248.
7. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J.* 1979;121:1193-1254.
8. Woolf SH, Battista RN, Anderson GM, et al. Assessing the clinical effectiveness of preventative meneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol.* 1990;43:891-905.
9. US Preventive Services Task Force. Screening for adolescent idiopathic scoliosis: review article. *JAMA.* 1993;269:2667-2672.
10. Hedges L, Olkin I. *Statistical Methods for Meta-analysis.* New York, NY: Academic Press Inc; 1985.
11. Diamond GA, Denton TA. Alternative perspectives on the biased foundations of medical technology assessment. *Ann Intern Med.* 1993;118:455-464.
12. Jackson R, Barham P, Bills J, et al. Management of raised blood pressure in New Zealand: a discussion document. *BMJ.* 1993;307:107-110.
13. Smith GD, Egger M. Who benefits from medical interventions? *BMJ.* 1993;308:72-74.
14. Glasziou P, Irwig L. Generalizing the results of clinical trials. Presented at the Second Cochrane Colloquium; October 2, 1994; Hamilton, Ontario.
15. Lubsen J, Tijssen JGP. Large trials with simple protocols: indications and contraindications. *Control Clin Trials.* 1989;10:151S-160S.
16. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1986;89(suppl 2):2S-3S.
17. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1992;102(suppl 4):305S-311S.
18. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to get started. *JAMA.* 1993;270:2093-2095.
19. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol.* 1994;47:881-889.
20. Jaeschke R, Guyatt GH, Shannon H, et al. Basic statistics for clinicians, III: assessing the effects of treatment: measures of association. *Can Med Assoc J.* 1995;152:351-357.
21. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA.* 1991;266:93-98.
22. Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med.* 1992;116:78-84.
23. Sacks HS, Chalmers TC, Smith H Jr. Sensitivity and specificity of clinical trials: randomized v historical controls. *Arch Intern Med.* 1983;143:753-755.
24. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358-1361.
25. Cook DJ, Reeve BK, Guyatt GH, Griffith LE, Heyland DK, Tryba M. Stress ulcer prophylaxis in the critically ill: resolving discordant meta-analyses. *JAMA.* In press.
26. Cook DJ, Fuller HD, Guyatt GH, et al. Risk factors for gastrointestinal bleeding in critically ill patients. *N Engl J Med.* 1994;330:377-381.
27. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med.* 1988;318:1728-1733.
28. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
29. Dobbs JH, Muir JG, Smith RN. H2-antagonists and hepatitis. *Ann Intern Med.* 1986;105:803.
30. Vial T, Goubier C, Begeret A, et al. Side effects of ranitidine. *Drug Saf.* 1991;6:94-117.
31. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life: basic sciences review. *Ann Intern Med.* 1993;70:225-230.
32. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ.* 1986;5:1-30.
33. Guyatt G, Jaeschke R, Cook DJ, Shannon H, Heddle N, Walter S. Basic statistics for clinicians, 2: interpreting study results: confidence intervals. *Can Med Assoc J.* 1995;152:169-173.
34. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: British Medical Journal; 1989:83-100.