Grading Strength of Recommendations and Quality of Evidence in Clinical Guidelines: Report From an American College of Chest Physicians Task Force

Gordon Guyatt, David Gutterman, Michael H. Baumann, Doreen Addrizzo-Harris, Elaine M. Hylek, Barbara Phillips, Gary Raskob, Sandra Zelman Lewis and Holger Schünemann

Chest 2006;129;174-181
DOI: 10.1378/chest.129.1.174

ISSN: 0012-3692.

CHEST

Special Features

Grading Strength of Recommendations and Quality of Evidence in Clinical Guidelines*

Report From an American College of Chest Physicians Task Force

Gordon Guyatt, MD, MSc, FCCP; David Gutterman, MD, FCCP; Michael H. Baumann, MD, MSc, FCCP; Doreen Addrizzo-Harris, MD, FCCP; Elaine M. Hylek, MD, MPH; Barbara Phillips, MD, FCCP; Gary Raskob, PhD; Sandra Zelman Lewis, PhD; and Holger Schünemann, MD, PhD, FCCP

While grading the strength of recommendations and the quality of underlying evidence enhances the usefulness of clinical guidelines, the profusion of guideline grading systems undermines the value of the grading exercise. An American College of Chest Physicians (ACCP) task force formulated the criteria for a grading system to be utilized in all ACCP guidelines that included simplicity and transparency, explicitness of methodology, and consistency with current methodological approaches to the grading process. The working group examined currently available systems, and ultimately modified an approach formulated by the international GRADE group. The grading scheme classifies recommendations as strong (grade 1) or weak (grade 2), according to the balance among benefits, risks, burdens, and possibly cost, and the degree of confidence in estimates of benefits, risks, and burdens. The system classifies quality of evidence as high (grade A), moderate (grade B), or low (grade C) according to factors that include the study design, the consistency of the results, and the directness of the evidence. For all future ACCP guidelines, The College has adopted a simple, transparent approach to grading recommendations that is consistent with current developments in the field. The trend toward uniformity of approaches to grading will enhance the usefulness of practice guidelines for clinicians. ***(CHEST 2006; 129:174–181)***

Key words: grading recommendations; grading system; methodology

Abbreviations: ACCP = American College of Chest Physicians; RCT = randomized controlled trial; RRR = relative risk reduction

Treatment decisions involve a tradeoff between benefits on the one hand, and risks, burdens, and, potentially, costs on the other. Guideline panels provide recommendations for the management of typical patients. To integrate these recommendations with their own clinical judgment, and with individual patient values and preferences, clinicians need to understand the basis for the recommendations that expert guidelines offer. A systematic approach to grading the strength of management recommendations can minimize bias and aid interpretation.[3] Indeed, most guideline groups have accepted the necessity for some sort of grading scheme.

While the grading of recommendations represents

*From the Departments of Medicine, and Clinical Epidemiology and Biostatistics (Dr. Guyatt), McMaster University, Hamilton, ON, Canada; the Department of Medicine (Dr. Gutterman), Medical College of Wisconsin, Milwaukee, WI; the University of Mississippi Medical Center (Dr. Baumann), Jackson, MS; New York University School of Medicine (Dr. Addrizzo-Harris), New York, NY; the Department of Medicine (Dr. Hylek), Research Unit-Section of General Internal Medicine, Boston University School of Medicine, Boston, MA; the University of Kentucky College of Medicine (Dr. Phillips), Lexington, KY; College of Public Health (Dr. Raskob), University of Oklahoma Health Sciences Center, Oklahoma City, OK; the American College of Chest Physicians (Dr. Lewis), Northbrook, IL; and the Department of Clinical Epidemiology, Italian National Cancer Institute (Dr. Schünemann), Rome, Italy.

Manuscript received August 16, 2005; revision accepted August 21, 2005.

Correspondence to: Gordon Guyatt, MD, MSc, FCCP, Department of Clinical Epidemiology and Biostatistics, HSC-2C12, McMaster University, 1200 Main St West, Hamilton, ON, Canada L8N 3Z5; e-mail: guyatt@mcmaster.ca

a positive development for guideline development and interpretation, the proliferation of grading systems has proved to be an unfortunate consequence. Methodologists and guideline developers have given much thought and effort to considering the criteria and approaches to an optimal grading system. The American College of Chest Physicians (ACCP) convened a working group to review the issue and to agree on a grading system that would be consistent with the latest developments in the field.

For editorial comment see pages 7 and 10

The task force began by developing criteria that define an optimal grading system (Table 1), placing them in an order that approximates their relative importance. These criteria guided the decisions of the group in the choice of the grading system that follows.

Strength of Recommendation

Guideline panels should make recommendations to administer, or not administer, an intervention, on the basis of tradeoffs between benefits on the one hand, and risks, burdens, and, potentially, costs on the other. If benefits outweigh risks and burdens, experts will recommend that clinicians offer a treatment to appropriately chosen patients. The uncertainty associated with the tradeoff between the benefits and the risks and burdens will determine the strength of recommendations.

The ACCP task force chose to classify recommendations into two levels, strong and weak (Table 2). If guideline panelists are very certain that benefits do, or do not, outweigh risks and burdens, they will make a strong recommendation, grade 1. If they think that the benefits and the risks and burdens are finely balanced, or if appreciable uncertainty exists about the magnitude of the benefits and risks, they must offer a weak, grade 2 recommendation.

A two-level grading system has the merit of simplicity. Two levels also facilitate the clear interpretation of the implications of strong and weak recommendations by clinicians. We offer three ways that clinicians can interpret strong and weak recommendations. We have already presented the first way. A strong recommendation signifies that benefits clearly outweigh the risks, or the reverse; a weak recommendation signifies that benefits and risks are closely balanced, or uncertain.

Clinicians are becoming increasingly aware of the importance of patient values and preferences in clinical decision making.[4] A second way to interpret strong and weak recommendations is in relation to patient values and preferences. For decisions in which it is clear that benefits far outweigh risks, or risks far outweigh benefits, virtually all patients will make the same choice (see box 1 for an example). In such instances, guideline panels can offer a strong (grade 1) recommendation. In contrast, there are other choices in which patient values and preferences will play a crucial role and in which patients will, as a result, make different choices. See boxes 2

Table 1—*Criteria for an Optimal Grading System*

Criteria	Description
1	Separation of grades of recommendations from quality of evidence
2	Simplicity and transparency for clinician consumer
3	Sufficient (but not too many) categories
4	Explicitness of methodology for guideline developers
5	Simplicity for guideline developers
6	Consistent with general trends in grading systems
7	Explicit approach to different levels of evidence for different outcomes

Box 1: Short-term aspirin reduces the relative risk of death after myocardial infarction by approximately 25%. Aspirin has minimal side effects and very low cost. Peoples' values and preferences are such that virtually all patients suffering a myocardial infarction would, if they understood the choice they were making, opt to receive aspirin. Guideline panels can thus offer a strong recommendation for aspirin administration in this setting.

Box 2: Consider a patient a 40 year-old man who has suffered an idiopathic deep venous thrombosis and has been taking adjusted dose warfarin for one year. If the patient continues on standard-intensity warfarin his risk of recurrent DVT will be reduced by approximately 10% per year.[1] The inevitable burdens of the treatment include taking a warfarin pill daily, keeping dietary intake of vitamin K constant, monitoring the intensity of anticoagulation with blood tests, and living with the increased risk of both minor and major bleeding. Some patients who are very averse to a recurrent DVT may consider the down sides of taking warfarin well worth it. Others are likely to consider the benefit not worth the risks and inconvenience.

Table 2—*Grading Recommendations*

Grade of Recommendation/ Description	Benefit vs Risk and Burdens	Methodological Quality of Supporting Evidence	Implications
1A/strong recommendation, high-quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	RCTs without important limitations or overwhelming evidence from observational studies	Strong recommendation, can apply to most patients in most circumstances without reservation
1B/strong recommendation, moderate quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	RCTs with important limitations (inconsistent results, methodological flaws, indirect, or imprecise) or exceptionally strong evidence from observational studies	Strong recommendation, can apply to most patients in most circumstances without reservation
1C/strong recommendation, low-quality or very low-quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	Observational studies or case series	Strong recommendation but may change when higher quality evidence becomes available
2A/weak recommendation, high-quality evidence	Benefits closely balanced with risks and burden	RCTs without important limitations or overwhelming evidence from observational studies	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2B/weak recommendation, moderate-quality evidence	Benefits closely balanced with risks and burden	RCTs with important limitations (inconsistent results, methodological flaws, indirect, or imprecise) or exceptionally strong evidence from observational studies	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2C/weak recommendation, low-quality or very low-quality evidence	Uncertainty in the estimates of benefits, risks, and burden; benefits, risk, and burden may be closely balanced	Observational studies or case series	Very weak recommendations; other alternatives may be equally reasonable

Box 3: A systematic review of randomized trials suggests that in 1,000 patients with ST elevation myocardial infarction who are receiving thrombolytic therapy and aspirin and who are treated with heparin (versus no treatment with heparin) 5 fewer will die, 3 fewer will have reinfarction, and 1 fewer will have a pulmonary embolus, while 3 more will have major bleeds.[2] Further, these estimates are not precise, and the advantage in decreased infarctions may be lost after six months. The small, imprecise and possibly transient benefit leaves us less confident about any recommendation to use heparin in this situation. Hence, the recommendation is likely to be weak.

and 3 for examples. When, across the range of patient values, fully informed patients are liable to make different choices, guideline panels should offer weak (grade 2) recommendations.

Following closely from this reasoning, a third way for clinicians to interpret strong recommendations is, for typical patients, to just do it. On the other hand, when clinicians face weak recommendations, or when they face patients with very atypical circumstances or values, they should carefully consider the benefits, risks, and burdens in the context of the individual patient before them.

How to individualize decision making in weak recommendations remains a challenge. One strategy uses decision aids that present patients with both the benefits and downsides of therapy.[5] Because of time constraints, clinicians cannot use decision aids in all patients. For strong recommendations, using a decision aid is likely, for most patients, to constitute a poor use of time and energy. For weak recommendations, clinicians should consider the use of a decision aid or, alternatively, a detailed conversation with the patient to ensure that the ultimate decision is consistent with the patient's values.

Factors That Influence the Strength of a Recommendation

Guideline panels must consider a number of factors in grading recommendations (Table 3). One issue is their confidence in the best estimates of benefit and harm. The rating of methodological quality, which we discuss below, captures that degree of confidence.

Table 3—*Factors Panels Should Consider in Deciding on a Strong or Weak Recommendation

Issue	Example
Methodological quality of the evidence supporting estimates of likely benefit, and likely risk, inconvenience, and costs	Many high-quality randomized trials have demonstrated the benefit of therapy with inhaled steroids in patients with asthma, while only case series have examined the utility of pleurodesis in patients with pneumothorax
Importance of the outcome that treatment prevents	Preventing postphlebitic syndrome with thrombolytic therapy in DVT patients in contrast to preventing death from PE
Magnitude of treatment effect	Clopidogrel vs aspirin leads to a smaller stroke reduction in patients with TIAs (RRR,[19] 8.7%) than anticoagulation vs placebo in patients with AF (RRR, 68%)
Precision of estimate of treatment effect	ASA therapy vs placebo in AF patients has a wider confidence interval than ASA therapy for stroke prevention in patients with TIA
Risks associated with therapy	ASA and clopidogrel for anticoagulation therapy in patients with acute coronary syndromes has a higher risk for bleeding than ASA alone
Burdens of therapy	Therapy with adjusted-dose warfarin is associated with a higher burden than that with aspirin; warfarin requires monitoring the intensity of anticoagulation and a relatively constant dietary vitamin K intake
Risk of target event	Some surgical patients are at very low risk of post-operative DVT and PE while other surgical patients have considerably higher rates of DVT and PE
Costs	Clopidogrel has a much higher cost in patients with TIA than does aspirin
Varying values	Most young, healthy people will put a high value on prolonging their lives (and thus incur suffering to do so); the elderly and infirm are likely to vary in the value they place on prolonging their lives (and may vary in the suffering they are ready to experience to do so)

*DVT = deep vein thrombosis; PE = pulmonary embolism; TIA = transient ischemic attack; AF = atrial fibrillation; ASA = aspirin.

The prevention of outcomes with high patient importance[6] should, in general, lead to stronger recommendations than the prevention of outcomes of lesser patient importance. For instance, one needs to expose four patients to a respiratory rehabilitation program for one patient to gain a small but important improvement in dyspnea in daily life.[7] In low-risk patients who have experienced a myocardial infarction, one might need to treat 100 patients with agents such as aspirin, β-blockers, angiotensin-converting enzyme inhibitors, or statins, to extend the life of one patient. Despite the much higher number needed to treat, since we value prolongation of life more highly than relieving dyspnea, the latter intervention may warrant a stronger recommendation.

The choice of adjusted-dose warfarin vs aspirin for the prevention of stroke in patients with atrial fibrillation illustrates a number of the factors that will influence the strength of a recommendation. A systematic review and metaanalysis[8] found a relative risk reduction (RRR) of 46% in all strokes with warfarin vs aspirin. This large effect supports a strong recommendation for warfarin. Furthermore, the relatively narrow 95% confidence interval (RRR, 29 to 57%) suggests that warfarin provides an RRR of at least 29%, and further supports a strong recommendation. At the same time, warfarin is associated with the inevitable burdens of keeping the dietary intake of vitamin K constant, monitoring the intensity of anticoagulation with blood tests, and living with the increased risk of both minor and major bleeding. Most patients, however, are much more stroke averse than they are bleeding averse.[8] As a result, almost all patients with high risk of stroke would choose therapy with warfarin, suggesting the appropriateness of a strong recommendation.

This last point emphasizes the importance of the patient's baseline risk of the adverse outcome that treatment is designed to avoid. Consider a 65-year-old patient with atrial fibrillation and no other risk factors for stroke. This individual's risk for stroke in the next year is approximately 2%. Therapy with dose-adjusted warfarin can, relative to aspirin, reduce the risk to approximately 1%. Some patients who are very stroke-averse may consider the downside of receiving warfarin therapy to be well worth it. Others are likely to consider the benefit not worth the risks and inconvenience. When, across the range of patient values, fully informed patients are liable to make different choices, guideline panels should offer weak (grade 2) recommendations.

As benefits and risks become more finely balanced, or more uncertain, decisions to administer an effective therapy also become more sensitive to resource use (cost) implications. When dealing with resource allocation issues, guideline panels face challenges of limited expertise, paucity of rigorous and unbiased cost-effectiveness analyses, and wide variability of costs across jurisdictions or health-care systems. Ignoring the issue of resource use (costs) is, however, becoming less and less tenable for guideline panels.[9]

When guideline developers make recommendations, they assume a particular set of values as they

weigh the possible beneficial and detrimental outcomes. When value or preference judgments are particularly salient, guideline panels should describe the key values attached to these outcomes and that influenced the direction of a recommendation or its grade. Guideline panels often do not elicit direct or indirect representation from patients in arriving at these values. Moreover, recommendations can only reflect average values. These considerations emphasize the importance of guideline panels making explicit the key values and preference judgments that drive their recommendations.

Wording of Recommendations

Given the proliferation of grading systems, and the resulting confusion, it is desirable to provide clinicians with as many indicators as possible in interpreting the strength of recommendations. ACCP panels, when they are making a strong recommendation, will use the terminology, "We recommend. . . . " When they make a weak recommendation, ACCP guideline panels will use less definitive wording, such as, "We suggest. . . . " Further, the clarity of recommendations requires that the target patient population be defined and, when appropriate, the details of how clinicians should administer the intervention.

Confidence in Estimates of Magnitude of Benefits, Risks, Burdens, and Costs

Early systems of grading methodological quality relied primarily on the basic study design (*ie*, randomized control trials [RCTs], or observational studies). The fundamental study design remains critically important in determining our confidence in estimates of beneficial and detrimental treatment effects. Because of prognostic differences between groups, and the lack of safeguards such as blinding that can avoid biased ascertainment of outcomes, evidence based on observational studies will, in general, be appreciably weaker than evidence from RCTs. The last several years have seen, however, an increased awareness of a number of other factors that influence our confidence in our estimates of risk and benefit (Table 4).

ACCP recommendations will henceforth use a three-category system of quality of evidence, as follows: high (grade A); moderate (grade B); and low quality (grade C) [Table 2]. Ideally, guideline panels will have available to them systematic reviews of the evidence regarding the benefits and risks of the alternative management strategies they are considering. Guideline panels will have the strongest evidence possible when such reviews reveal one or more well-designed and well-executed RCTs yielding consistent directly applicable results. Strong evidence can also come, under unusual circumstances, from observational studies yielding very large effects.

Table 4—*Factors Panels Should Consider in Deciding on Their Confidence in Estimates of Benefits, Risks, Burden, and Costs*

Factor Type	Factors
Factors that may decrease the quality of evidence based on RCTs	Poor quality of planning and implementation of the available RCTs suggesting high likelihood of bias
	Inconsistency of results
	Indirectness of evidence
	Sparse evidence
Factors that may increase the quality of evidence based on observational studies	Large magnitude of effect
	All plausible confounding would reduce a demonstrated effect
	Dose-response gradient

The moderate quality category is populated by randomized trials with important limitations and by exceptionally strong observational studies. Observational studies, and on occasion RCTs with multiple serious limitations, will fill the low-quality evidence category. This categorization follows the principle that all relevant clinical studies provide evidence, the quality of which varies. Following this principle, the ACCP does not use a threshold for "acceptable evidence" in the peer-reviewed published medical literature.

Factors That Modify the Quality of Evidence: Limitations in RCTs

When RCTs have addressed the impact of alternative management strategies (both benefits and harms) on all relevant outcomes, they will yield high-quality evidence unless they have one of a number of limitations. The following limitations may decrease the quality of evidence supporting a recommendation (Table 4).

1. Our confidence in recommendations decreases if the available RCTs have major deficiencies that are likely to result in a biased assessment of the treatment effect. These methodological limitations include a very large loss to follow-up, or an unblinded study with subjective outcomes that are highly susceptible to bias. How lack of blinding can influence the grading is exemplified by a recommendation to treat heparin-induced thrombocytopenia complicated by thrombosis with danaparoid sodium. The

randomized trial evidence for danaparoid use in patients with heparin-induced thrombocytopenia comes from an unblinded trial[20] in which the outcome was the clinicians' assessment of when the thromboembolism had resolved, which is a subjective judgment. As a result, an ACCP guideline panel rated the quality of the evidence as moderate rather than strong.[10]

2. When several RCTs yield widely differing estimates of treatment effect (heterogeneity or variability in results) investigators look for explanations for that heterogeneity. For instance, drugs may have larger relative effects in sicker, or in less sick, populations. When heterogeneity exists, but investigators fail to identify a plausible explanation, the strength of recommendations from even rigorous RCTs is weaker. For example, RCTs of pentoxifylline in patients with intermittent claudication have shown conflicting results that so far defy explanation. Acknowledging the unexplained heterogeneity, an ACCP guideline panel rated the quality of the evidence for pentoxifylline as moderate, rather than high.[11]
3. Investigators may have undertaken RCTs in similar populations, but not identical populations, to those of interest to a guideline panel. Panels should consider this to be indirect evidence and, to the extent they are uncertain about the applicability to their relevant population, should downgrade the quality of evidence. For instance, while graduated compression stockings have proven to be of benefit in a variety of populations at risk for venous thrombosis, they have never been tested directly in trauma patients. An ACCP guideline panel judged the available RCTs to be relevant to trauma patients in whom the administration of low-molecular-weight heparin is contraindicated, but because of concern about generalizing from other populations (that is, concern about the indirectness of the evidence), rated the quality of the evidence as moderate. Had they had no concerns about directness, they would have considered the evidence to be of high quality, whereas if there were no relevant RCTs available, and the best evidence came from observational studies, they would have rated the evidence to be of low quality.[12] Indirectness may also apply to the intervention [(*eg*, RCTs of similar but not identical interventions or different doses and formulations)] and outcomes (*eg*, RCTs measuring laboratory exercise capacity when a panel is really interested in quality-of-life improvement).
4. Investigators may have conducted RCTs, but included very few patients and observed very few events. For instance, a well-designed and rigorously conducted RCT addressed the use of nadroparin, a low-molecular-weight heparin, in patients with cerebral venous sinus thrombosis. Of 30 treated patients, 3 had a poor outcome, as did 6 of 29 patients in the control group. The investigators' analysis suggested a 38% reduction in the relative risk of a poor outcome, but the result was not statistically significant.[13] Because of the small number of patients, and the small number of events, an ACCP guideline panel judged the quality of the evidence for anticoagulation in cerebral sinus thrombosis as moderate rather than high.[12]

Factors That Modify the Quality of Evidence: Observational Studies Can Provide Moderate or Strong Evidence

While observational studies will generally yield only low-quality evidence, there may be unusual circumstances in which guideline panels will classify such evidence as of moderate quality, or even high quality.

1. On the rare occasions when they yield extremely large and consistent estimates of the magnitude of a treatment effect, we may be confident about the results of observational studies. For example, oral anticoagulation in mechanical heart valves has not been compared to placebo in an RCT. However, evidence from observational studies suggests that the probability of experiencing thromboembolic events without anticoagulation is 12.3% annually in patients with bileaflet prosthetic aortic valves and higher for those with other valve types,[14] and estimates of the RRR with oral anticoagulation are in the range of 80%. While the observational studies are likely to overestimate the true effect, the weak study design is very unlikely to explain the entire benefit. Thus, an ACCP guideline panel concluded that these data, despite the absence of randomized trials, constituted strong evidence of the effectiveness of anticoagulation in bileaflet aortic prosthetic valves.[15]
2. On equally rare occasions, all plausible biases from observational studies may be working to underestimate an apparent treatment effect. In other words, the actual treatment effect is very likely to be larger than what the data suggest. For instance, a rigorous systematic review of observational studies including a total of 38 million patients compared private for-profit vs

private not-for-profit hospital care. The meta-analysis[16] demonstrated higher death rates in the private for-profit hospitals.

The investigators postulated two likely sources of bias. The first was residual confounding with disease severity. It is likely that, if anything, patients in the not-for-profit hospitals were sicker than those in the for-profit hospitals. Thus, to the extent that residual confounding existed, it would bias results against the not-for-profit hospitals.

The second likely bias was the possibility that higher numbers of patients with excellent private insurance coverage could lead to a hospital having more resources and to a "spillover" effect that would benefit those without such coverage. Since for-profit hospitals are likely to admit a larger proportion of such well-insured patients than are not-for-profit hospitals, the bias is once again against the not-for-profit hospitals. Because the plausible biases would all diminish the demonstrated treatment effect, one might consider the evidence from these observational studies as being of moderate quality rather than of low quality.

What to Do When Quality of Evidence Differs Across Outcomes?

When RCT results are available, the quality of evidence will often differ between primary efficacy and toxicity outcomes, usually between efficacy outcomes and cost, and almost always between efficacy outcomes and rare but serious side effects. On most occasions, efficacy outcomes will be the most important, and guideline panels can base their rating of the quality of the evidence exclusively on these end points. Panels should, however, consider whether toxicity end points are also crucial to the decision regarding the optimal management strategy. If they are, panels should consider the quality of evidence regarding those end points, and should make a final rating about the quality of evidence accordingly.

For instance, consider a guideline panel addressing the use of long-term oral steroids for patients with stage 2 or 3 sarcoidosis with moderate-to-severe symptoms and radiographic changes. Randomized trials have addressed the impact of steroids on radiographic findings, symptoms, and spirometry over a period of 2 years.[17] These trials failed, however, to address steroid toxicity. If a guideline panel ignored toxicity, they might well rate the quality of evidence as high. If, however, they consider steroid toxicity as crucial in their decision, the uncertainty about the impact of treatment increases. If they look for observational studies to estimate steroid toxicity, the quality of the evidence about toxicity is likely to be low, and this may be the most appropriate rating for the overall quality of evidence. Alternatively, they may seek randomized trials of steroids in other conditions and face limitations of directness. They may then conclude that the evidence regarding steroid toxicity, and the overall quality of the evidence, is moderate.

The ACCP Grading System and Initiatives Toward Uniform Grading Across Guideline Panels

In considering alternative grading systems, we found that the structure and guides for application and interpretation suggested by the GRADE group largely met the criteria in Table 1.[18] As a result, the categories presented in Table 2 permit similar interpretation to those of the GRADE group. The important aspect in which the ACCP task force approach differs is in combining low-quality and very low-quality evidence. While we achieved the primary goal of the ACCP task force, to identify a unified grading system for all future ACCP evidence-based guidelines, this exercise went beyond that goal. This article will facilitate the adoption of uniform guidelines through a simple, straightforward presentation that any guideline panel interested in the principles underlying Table 2 will find useful.

Clinicians' understanding of systems of grading the strength of recommendations and quality of evidence will also benefit if systems map easily onto one another. The ACCP mapping onto the GRADE system is obvious, and the approach that the ACCP has adopted also maps easily onto other systems, including that of the ACC/AHA and prior ACCP guideline grading systems, further facilitating understanding and usefulness.

Summary

In the system that the ACCP has adopted, the strength of any recommendation depends on the following two factors: the tradeoff between the benefits and the risks and burdens; and the quality of the evidence regarding treatment effect. We grade the tradeoff between the benefits, and the risks and burdens into the following two categories; category 1, in which the tradeoff is clear enough that most patients, despite differences in values, would make the same choice, leading to a strong recommendation; and category 2, in which the tradeoff is less clear, and individual patient values will likely lead to different choices, leading to a weak recommendation. We grade methodological quality in terms of the following three categories: randomized trials that show consistent results, or observational studies with

very strong treatment effects; randomized trials with limitations, or observational studies with exceptional strengths; and observational studies without exceptional strengths and case series. The framework summarized in Table 2 generates recommendations from the very strong (benefit/risk tradeoff unequivocal, high-quality evidence, grade 1A) to the very weak (benefit/risk questionable, low-quality evidence, grade 2C). Whatever the grade of the recommendation, clinicians must use their judgment, considering both local and individual patient circumstances, and patient values, in making individual decisions. In general, however, they should place progressively greater weight on expert recommendations as they move from grade 2C to grade 1A.

References

1 Büller H, Agnelli G, Hull R, et al. Antithrombotic therapy for venous thromboembolic disease. Chest 2004; 126:401S–428S
2 Collins R, MacMahon S, Flather M, et al. Clinical effects of anticoagulant therapy in suspected acute myocardial infarction: systematic overview of randomised trials. BMJ 1996; 313:652–659
3 Guyatt G, Sinclair J, Cook D, et al. Grading recommendations: a qualitative approach. In: Guyatt GR, Rennie D, ed. Users' guide to the medical literature: a manual for evidence-based practice. Chicago, IL: AMA Press, 2002
4 Guyatt G, Straus S, McAlister F, et al. Incorporating patient values. In: Guyatt GR, Rennie D, ed. Users' guide to the medical literature: a manual for evidence-based practice. Chicago, IL: AMA Press, 2002
5 O'Connor A, Stacey D, Entwistle V, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev, Issue 2, (database online) 2003
6 Guyatt G, Montori V, Devereaux P, et al. Patients at the centre: in our practice, and in our use of language. ACP J Club 2004; 140:A11–A12
7 Goldstein R, Gort E, Guyatt G, et al. Economic analysis of respiratory rehabilitation. Chest 1997; 112:370–379
8 Devereaux P, Anderson D, Gardner M, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. BMJ 2001; 323:1218–1222
9 Guyatt G, Baumann M, Pauker S, et al. Addressing resource allocation issues in recommendations from clinical guideline panels. Chest 2006; 129:182–187
10 Warkentin T, Greinacher A. Heparin-induced thrombocytopenia: recognition, treatment, and prevention: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest 2004; 126:311S–337S
11 Clagett G, Jackson M, Lip G, et al. Antithrombotic therapy in peripheral arterial occlusive disease: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest 2004; 126:609S–626S
12 Geerts W, Pineo G, Heit J, et al. Prevention of venous thromboembolism: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest 2004; 126: 338S–400S
13 de Bruijn S, Stam J. Randomized, placebo-controlled trial of anticoagulant treatment with low-molecular-weight heparin for cerebral sinus thrombosis. Stroke 1999; 30:484–488
14 Baudet E, Puel V, McBride J, et al. Long-term results of valve replacement with the St. Jude Medical prosthesis. J Thorac Cardiovasc Surg 1995; 109:858–870
15 Salem D, Stein P, Al-Ahmad A, et al. Antithrombotic therapy in valvular heart disease—native and prosthetic: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest 2004; 126:457S–482S
16 Devereaux P, Choi P, Lacchetti C, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ 2002; 166:1399–1406
17 Paramothayan S, Jones P. Corticosteroid therapy in pulmonary sarcoidosis: a systematic review. JAMA 2002; 287:1301–1307
18 Atkins D, Best D, Briss P, et al. Grading quality of evidence and strength of recommendations. BMJ 2004; 328:1490
19 CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). Lancet 1996; 348:1329–1339
20 Chong BH, Gallus AS, Cade JF, et al. Prospective randomised open-label comparison of danaparoid with dextran 70 in the treatment of heparin-induced thrombocytopenia with thrombosis: a clinical outcome study. Thromb Haemost 2001; 86:1170–1175

Grading Strength of Recommendations and Quality of Evidence in Clinical Guidelines: Report From an American College of Chest Physicians Task Force

Gordon Guyatt, David Gutterman, Michael H. Baumann, Doreen Addrizzo-Harris, Elaine M. Hylek, Barbara Phillips, Gary Raskob, Sandra Zelman Lewis and Holger Schünemann

Chest 2006;129;174-181
DOI: 10.1378/chest.129.1.174