



Misconceptions, Challenges, Uncertainty, and Progress in Guideline Recommendations

Regina Kunz,^a Benjamin Djulbegovic,^b Holger J. Schunemann,^c Martin Stanulla,^d Paula Muti,^c and Gordon Guyatt^e

Although the quality of clinical practice guidelines has improved over the last decade, current guideline systems have limitations that reduce their validity and limit their acceptance. The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) working group, a worldwide collaboration of guideline developers, methodologists, and clinicians, has constructed a system for developing guidelines that addresses these shortcomings. The system includes a transparent and rigorous methodology for rating the quality of evidence, an explicit balancing of benefits and harms of healthcare interventions, an explicit acknowledgement of the values and preferences that underlie the recommendations, and whether the intervention represents a wise use of resources. These four elements determine whether a recommendation is strong or weak. A guideline panel offers strong recommendations when virtually all informed patients would choose the same management strategy. Weak recommendations imply that choices will differ across the range of patient values and preferences. The GRADE system has been tested in multiple practice settings and for a large spectrum of questions, refined and re-evaluated to ensure that it captures the complex issues involved in evidence assessment and grading recommendations while maintaining simplicity and practicality. Many guideline organizations and medical societies have endorsed the system and adopted it for their guideline processes.
Semin Hematol 45:167-175 © 2008 Elsevier Inc. All rights reserved.

Guidelines have become popular tools to support clinicians, offering evidence-based management strategies to help clinicians deliver the best care for individual patients, and ultimately to improve health outcomes. Established guideline organizations and medical societies have invested considerable effort, methodological expertise, and resources to develop structured approaches to meet these goals. While these efforts have advanced the science of guideline development, the limitations of existing systems have become increasingly evident.¹⁻⁴ These limitations provided the impetus for GRADE (Grades of Recommendation, Assessment,

Development and Evaluation),⁵ an initiative of a widely representative group of international guideline developers, methodologists, and clinicians who set out to create a comprehensive framework for developing guidelines based on the following principles:

- a clear distinction between quality of evidence and strength of recommendations
- recognition of the diversity of patient-important outcomes influenced by alternative management strategies
- acknowledgement of the values and preferences underlying specific recommendations and management decisions
- the integration of resources use in the balance of desirable and undesirable consequences of health care interventions
- a transparent process to move from evidence to recommendations
- comprehensiveness and simplicity.

In this article, we describe the GRADE system starting with graded recommendations and how to get there, followed by a more detailed discussion of the quality assessment of the

^aBasel Institute of Clinical Epidemiology, Basel, Switzerland.

^bH. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, FL.

^cItalian National Cancer Institute "Regina Elena," Rome, Italy.

^dDepartment of Pediatric Hematology and Oncology, Medical School of Hannover, Hannover, Germany.

^eDepartment of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.

Dr Kunz has been supported by Santesuisse and the Gottfried and Julia Bangerter-Rhyner Foundation.

Address correspondence to Regina Kunz, MD, MSc (Epi), Basel Institute for Clinical Epidemiology, University Hospital Basel, Hebelstrasse 19, CH-4031 Basel, Switzerland. E-mail: rkunz@uhbs.ch

evidence focusing on the perspective of the hematologist/hemato-oncologist clinician.

Steps in the Development of Recommendations in the GRADE System

Using the GRADE approach to developing guidelines involves two steps: first, consideration of the quality of the evidence; second, development of a graded recommendation that considers not only the quality of the evidence but also the balance of desirable and undesirable consequences of an intervention, the (un-)certainty regarding the values and preferences patients attribute to the desirable and undesirable consequences, and resource use.

Strength of Recommendations: How to Grade and Why

The GRADE system offers two categories of recommendations: strong and weak, either of which can be for, or against, an intervention in comparison to an alternative. For GRADE, the strength of a recommendation describes to what extent guideline developers are confident that the desirable consequences of an intervention outweigh its undesirable consequences. Desirable interventions reduce morbidity and mortality, improve quality of life, decrease inconvenience, and incur few adverse effects or major resource use (cost). In making a strong recommendation, a guideline panel expresses its confidence that the desirable consequences of adhering to a treatment will outweigh any undesirable consequences. In making a weak recommendation, a guideline panel communicates a less confident judgment: the desirable consequences of adhering to a recommendation will probably outweigh the undesirable consequences.

Many guideline panels classify their recommendations into three, four, or more categories, sometimes including a separate category for expert opinion. Why did GRADE confine itself to only two? Driven by its goal of simplicity, the GRADE group perceived a great difficulty in translating nuances in uncertainty into meaningful differences in practical advice for clinicians and patients. Furthermore, when faced with inevitable shades of gray, a system with three or more categories runs the risk of a large proportion of recommendations in the intermediate category, leaving clinicians with little benefit from the grading exercise. The GRADE system eschews the category “expert opinion” because it recognizes that expert judgment is always needed for interpreting evidence, and because some evidence—albeit unsystematic clinical observations or indirect evidence from animal or cellular research—is almost always available. The two categories of recommendations of the GRADE system, strong and weak, help achieve the objective of simplicity and lead to clear directives for action that we shall describe shortly.

Guidelines for Whom?

In the past, clinician experts wrote guidelines primarily for their peers to help them select the best treatment for their patients. The increasing access of patients to healthcare information, together with the movement for consumer and patient autonomy, has made patients a receptive audience for structured healthcare information and advice, thus creating a partnership with health care professionals. A final audience consists of those involved in regulation: guidelines have become a source for “standards of care” to assess the performance of individual clinicians.

Interpreting Strong and Weak Recommendations

As defined above, a strong recommendation indicates that the desirable consequences of an intervention will very likely outweigh any potential undesirable consequences, while a weak recommendation leaves that judgment less certain. The implications of a recommendation, whether strong or weak, will vary depending on the target audience. For patients, a strong recommendation implies that, when fully informed, they are very likely to make the same choice. Use of a decision aid—a tool to inform patients and help them to identify their preferred choice among two or more treatment options^{6,7}—is unlikely to be efficient, since almost all patients would choose the same option.

For clinicians, a strong recommendation indicates that almost all patients should receive the intervention and that they should advise their patients accordingly. Healthcare policy-makers can draw the conclusion that this intervention may be a suitable candidate for a standard of care criterion or quality indicator.

In contrast, weak recommendations tell patients that the majority would want the recommended course of action but that a considerable proportion would opt for an alternative. Here, decision aids are likely to help patients choose the option that is most consistent with their individual values and preferences. For clinicians, weak recommendations indicate that patients are likely to differ in their preferences and choices and they should assist patients in finding their preferred treatment option. Healthcare policy-makers should seldom if ever choose a weak recommendation as a quality criterion. However, taking action to ensure patients are fully informed, and their values and preferences considered in the decision, may be a quality indicator.

Factors Affecting the Strength of a Recommendation

While many organizations agree that high-quality evidence need not result in a strong recommendation,³ few have made explicit the process of translating evidence into recommendations. The GRADE system has identified four factors that determine the direction and strength of a recommendation (Table 1).

Table 1 Factors That Affect the Strength of a Recommendation—Further Examples

Factor	Examples of Strong Recommendations*	Examples of Weak Recommendations†
1. Quality of evidence	Adjuvant chemotherapy for early breast cancer improves survival and decreases the rate of breast cancer recurrence. ¹⁷	Only case series have examined the utility of immunosuppressive therapy for treatment of erythropoietin-induced pure red cell aplasia; 37 of 47 patients (78%) who received immunosuppressive therapy recovered after treatment with corticosteroids, corticosteroids plus cyclophosphamide, or cyclosporine. ¹⁸
2. Uncertainty about the balance between beneficial and harmful consequences	In patients with myeloma and bone involvement, bisphosphonates reduce skeletal-related morbidity, at low toxicity, inconvenience, and moderate cost. ¹⁹	Granulocyte colony-stimulating factor in patients with cancer and chemotherapy-induced febrile neutropenia shortens hospital stay, but causes significant bone pain, joint pain, and flu-like symptoms. ²⁰
3. Uncertainty or variability in values and preferences	Almost all patients with large-cell lymphoma will place a higher value on the life-prolonging effects of chemotherapy over treatment toxicity.	In choosing allogeneic versus autologous stem cell transplant/chemotherapy for acute myelogenous leukemia, patients may or may not be willing to accept higher treatment-related mortality associated with allogeneic transplant (21% v 4%) even though allogeneic transplant improves leukemia-free survival (48% v 37% at 4 years). ²¹
4. Uncertainty about whether the intervention represents a wise use of resources	In patients with polycythemia vera, low-dose aspirin effectively reduces the risk of common complications such as nonfatal myocardial infarction, nonfatal stroke, pulmonary embolism, major venous thrombosis, or death from cardiovascular causes (RR, 0.40; 95% CI, 0.18 to 0.91), at no increase in bleeding complications, minimal inconvenience and cost. ²²	Although rasburicase may be superior to allopurinol for prevention of tumour lysis syndrome, its high treatment cost reduces the likelihood of a strong recommendation for its use. ²³

*Relates to cases when the quality of evidence is high and uncertainty about other factors is low.

†Relates to cases when the quality of evidence is low and uncertainty about other factors is high.

Quality of Evidence

Guideline panels are more likely to offer strong recommendations if the evidence comes from studies with robust methodology. For example, evidence from 27 rigorous randomized controlled trials (RCTs), including more than 6,000 patients that were pooled in an individual patient data (IPD) meta-analysis, demonstrated that more toxic combination chemotherapy with three and more drugs for patients with multiple myeloma failed to achieve superior survival compared to less toxic treatment with melphalan plus prednisone (odds ratio, 0.98; 95% confidence interval [CI], 0.92 to 1.04).⁸ This direct, consistent evidence, with a tight confidence interval and little risk of publication bias, leaves us with great confidence in the estimate of effect and constitutes high-quality evidence.

Evidence, however, is much weaker for patients with myeloma and pancytopenia. High-dose dexamethasone might be an option as initial therapy, but the benefit for this treatment is documented in only one observational study⁹ without a direct comparison group, including few patients, and failing to achieve complete follow-up. We are much less con-

fident in this estimate of effect, and the evidence for the benefit of high-dose dexamethasone relative to supportive care on survival is therefore very low quality.

Balancing Desirable and Undesirable Consequences

Desirable Consequences Clearly Outweigh Undesirable Consequences

Consider a 40-year-old man who has suffered an idiopathic deep venous thrombosis (DVT). The initial decision to use, for 6 months to a year, anticoagulants is not difficult: the desirable consequences of preventing recurrent DVT and pulmonary embolism clearly outweigh the bleeding risk.¹⁰ For such patients, initial use of anticoagulants warrants a strong recommendation.

On occasion, faced with only low-quality evidence, authors may nevertheless make strong recommendations. For instance, consider the recommendation for routine monitoring of platelet counts in patients receiving heparin whose risk of heparin-induced thrombocytopenia (HIT) is greater than

0.1%. Although the evidence of benefit is weak, the early discontinuation of heparin when the platelet count drops and other evidence to support HIT exists may be of appreciable benefit, and the costs and risks of monitoring are negligible. Thus, monitoring may warrant a strong recommendation.

Tight Balance Between Desirable and Undesirable Consequences

For the patient with venous thrombosis described above, continuing on standard-intensity warfarin beyond 1 year will reduce his absolute risk for recurrent DVT by more than 7% per year for several years.¹⁰ However, the burdens of treatment include taking warfarin 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. Patients who are very averse to a recurrent DVT would consider the benefits of avoiding DVT worth the downsides of taking warfarin. Other patients are likely to consider the benefit not worth the harms and burden. Thus, a weak recommendation for anticoagulation beyond 1 year is warranted.

(Un-)Certainty About Values and Preferences

A third determinant of the strength of a recommendation is the spectrum of values and preferences that might drive the patients' decisions. Values and preferences bear on the patients' reaction to the outcomes of the disease, the complications of the therapy, the inconveniences and burden of the therapy to patients and their family, or the cost associated with the treatment.

Consider the alternatives facing pregnant women with DVT requiring full anticoagulation. Standard drug treatment with warfarin puts the fetus at a low risk of relatively minor developmental abnormalities between the 6th and 12th week of pregnancy. The alternative, heparin, eliminates the risk to the child, but comes with disadvantages of pain (heparin injections), inconvenience, and cost. Nevertheless, obstetricians observe that women overwhelmingly place a very high value on preventing fetal complications. Here, a strong recommendation for heparin substitution is warranted.

Contrast this with the situation facing patients with multiple myeloma who need to decide about adjunct treatment with alpha-interferon. A systematic review and IPD meta-analysis of 24 RCTs confirmed a small advantage in survival for interferon (40 months with versus 36 months without alpha-interferon) but also demonstrated considerable toxicity (including severe flu-like symptoms, fever, and neurological symptoms) and the inconvenience of self-injection.¹¹ When interviewed regarding their preference for interferon therapy, half the patients felt the drug's toxicity an acceptable trade-off for a 6 months' gain in progression-free survival or overall survival, while around 30% did not, and the rest were undecided.¹² The interviews clearly documented how differently patients valued the small survival benefit relative to the undesirable consequences, thus justifying only a weak recommendation in favor of alpha-interferon.

Evidence concerning the values and preferences of patients is often unavailable, leaving guideline panels the alternatives to commission their own studies, to ask for support from consumer groups, or to speculate about the potential spectrum of patient values and preferences. Whatever approach a panel takes, for recommendations that are particularly value-sensitive, it should always explicitly inform the reader about the values that underlie their recommendations. For instance, in its guidelines on avian flu, the World Health Organization (WHO) noted that its recommendation to administer combination therapy with neuraminidase-inhibitor plus M2-inhibitor for patients with strongly suspected infection placed a high value on the prevention of death in an illness with a high case-fatality, and a relatively low value on adverse effects, drug resistance, and costs.¹³

Resource Implications

The final determinant of the strength of a recommendation is resource use (cost). Relative to other outcomes, cost is much less consistent over time, geographic areas, and implications than are other outcomes, and the value healthcare systems attach to the cost-benefit ratio differs considerably. Prices for the same drug differ widely across jurisdictions, as do resource implications, and recommendations that are heavily influenced by costs are likely to change over time as resource implications evolve. Furthermore, highly variable costs tend to make a recommendation very context-sensitive.

For example, prior to the recent warning of the US Food and Drug Administration¹⁴ regarding the safe use of erythropoiesis-stimulating agents (ESAs) in patients with cancer, the use of ESAs in cancer-associated anemia was thought to improve quality of life while the impact on survival remained uncertain. Some healthcare systems considered the moderate improvement in quality of life worth the cost,¹⁵ while others regarded the marginal improvement insufficient to justify the costly treatment.¹⁶

Guideline panels must explicitly specify the setting to which a recommendation applies and whose perspective they used when considering costs. All of these limitations make it less likely that cost will impact on direction or strength of a recommendation.

Rating Quality of Evidence: What Aspects to Consider?

The starting point for any quality assessment is a structured question that specifies the population, the intervention(s), the comparator(s), and the outcomes of interest, such as: "In lymphoma patients at risk of developing chemotherapy-induced febrile neutropenia, what are the benefits and harms associated with the use of granulocyte colony-stimulating factor (G-CSF) compared to not using G-CSF?"²⁴ Under ideal circumstances, a guideline panel can find the answers to its question in a recent systematic review and meta-analysis that has compiled the current knowledge and conducted an appropriate evaluation and analysis. In the absence of systematic reviews, the panel must do its best to identify and assess

Table 2 Hierarchy of Patient Important Outcomes for Decision-Making: Benefits and Harms of G-CSF in Cancer Patients at Risk for Febrile Neutropenia²⁴

	Benefits	Harms
Critical	Overall survival Reduction in infection-related mortality Reduction in progression or relapse of disease Reduction in infections	Splenic rupture Thromboembolic complications Bone pain
Important	Higher rates of complete response Shorter duration of hospitalization Reduction of stomatitis Reduction in antibiotic treatment Reduction in (febrile) neutropenia Shorter duration of high-intensity hygienic precautions	Asthenia (fatigue) Fever Rash Injection site reaction and pain
Less Important	Reduction in hospitalization cost (though some might call this outcome important)	Splenomegaly Peripheral edema (mild) Constipation (mild)

all the relevant best evidence, and clinicians must judge how well the panel has achieved this goal.

Clinicians should also attend to the outcomes a guideline panel has considered. Does the recommendation take into account factors such as adverse events, impact on quality of life, or long-term outcomes?^{25,26} Here, the GRADE system distinguishes outcomes that are critical to decision-making, outcomes important to decision-making and outcomes of limited importance. Critical to decision-making means that careful balance of those outcomes will determine the final decision, while “outcomes important to decision-making” will be taken into consideration but will not drive the decision. Table 2 lists such a hierarchy for the question of G-CSF use in lymphoma.

Rating Quality of Evidence— The Clinical Context Matters

The GRADE system classifies evidence for both beneficial and harmful effects into four different levels: high, moderate, low, and very low quality. Table 3 explains the implications of each level.

Table 3 The Definitions for Various Quality Levels of the Evidence

Level	Definition
High quality	Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality	Any estimate of effect is very uncertain.

The definitions reflect the extent to which we are confident that an estimate of effect is correct. Guideline panels need to judge the quality of evidence relative to the particular clinical context for which they are using the evidence. Therefore, in the context of guidelines, the quality of evidence reflects the extent to which our confidence in an estimate of the effect is adequate to support a particular recommendation.

For example, consider the choice of clopidogrel or aspirin in patients who have experienced a transient ischemic attack. A concealed, blinded randomized trial of more than 19,000 patients showed a relative risk reduction of 9% in ischemic events with clopidogrel over aspirin. Ordinarily, one would consider this extremely rigorous very large trial as high-quality evidence. Unfortunately, the confidence interval includes a relative risk reduction of less than 1%, which, if true, would not warrant administration of the much more expensive clopidogrel. Thus, the quality of the evidence is only moderate to support the decision for use of clopidogrel.

Quality of Evidence Has a Major Impact on the Grade of a Recommendation

Like many other rating systems, GRADE’s quality assessment starts (but does not end—see following) with the study design. Results from randomized controlled trials are regarded more trustworthy and robust than results from observational studies, for good reasons. Consider the optimal target level of hemoglobin in patients with renal anemia. Experts have, on the basis of observational studies in which patients with higher levels demonstrated superior outcomes to those with lower levels, advocated close to normal hemoglobin level as an optimal target in administering erythropoietin. However, when the ideal target level was tested in a RCT normalized hemoglobin levels resulted in higher mortality than lower levels.²⁷

The quality of the evidence depends not only on the study design and implementation²⁸ but also on consistency of results, directness of comparisons, precision of effect estimates, and likelihood of publication bias. Furthermore, large or very large treatment effects, a clear relationship

between dose and response, or studies where all plausible biases potentially decreased the magnitude of an effect can raise the quality of the evidence.

Downgrading the Evidence

Study Limitations

Important weaknesses in the planning, conduct, and analysis of studies reduce quality of evidence. Adjuvant chemotherapy complementing surgery is an option for invasive bladder cancer. A recent meta-analysis included 491 patients from six trials and showed a survival hazard ratio of 0.75 ($P = .019$), suggesting a reduction in the relative risk of death for chemotherapy versus control arms.²⁹ However, the analysis included trials that had stopped early for benefit³⁰ and trials seriously compromised by limitations in study design. These problems raise serious concern about the validity of the results²⁹ and justify downgrading the evidence to low quality.

Heterogeneity in Results

If the existing studies show inconsistent effects and this heterogeneity remains unexplained, as it often does, the quality of the evidence decreases. For example, 18 studies including 1,409 patients suggested that in cancer patients with chemotherapy-induced neutropenia quinolone reduced the risk of fever versus placebo or no intervention by 33% (relative risk [RR], 0.67 [0.56 to 0.81]). The results were compromised by great variability between studies (reflected in an I^2 -value of 86.2%) that could not be explained by any identifiable difference between studies (Table 4).³¹ Such heterogeneity would downgrade the quality of the evidence.

Indirect Evidence

If evidence is only indirectly related to the question of interest, guideline panels should downgrade the quality of evidence. A typical situation is the lack of head-to-head comparisons among drugs of the same class. Imagine that you would like to know whether pamidronate or clodronate is superior in lowering hypercalcemia and controlling pain in myeloma patients with bone involvement. All studies, however, compared clodronate to placebo and pamidronate to placebo, mandating a lower quality judgment on the basis of the lack of head-to-head comparisons.¹⁹ Indirectness also can refer to differences in the population, intervention, or outcomes of interest for the guideline versus those studied in the populations in trials. For example, comparisons of epoetin versus darbepoetin have been performed in patients with solid organ tumors but not in patients with hematological malignancies.³²

Imprecision

Studies including relatively few patients and/or observing only few events generally result in wide confidence intervals and considerable uncertainty about the true treatment effect. Consider a study of fluorochinolone prophylaxis in reducing infection-related mortality in neutropenic cancer patients that reported a relative risk of 0.38 (95% CI, 0.21 to 0.69). The relatively wide confidence interval does not fully capture the uncertainty associated with this estimate: it is based on a total of only 47 events. Moving only a few deaths

Table 4 Summary (GRADE) Evidence Profile for Quality Assessment and the Summary of Findings Table for the Meta-analysis "Antibiotic Prophylaxis Reduces Mortality in Neutropenic Patients: Fluorochinolones Versus Placebo or No Intervention"³¹

	Quality Assessment					Summary of Findings			Quality
	Limitations	Consistency	Directness	Precision	Publication Bias	Events/Participants	Relative Risk (95% CI)	No. of Participants (studies)	
Outcome: All cause mortality									
Downgrade -1*	No important inconsistency	No uncertainty	Imprecision	No publication bias	33/652 = 5.06	59/592 = 9.96	0.52 (0.35 to 0.77)	1,244 pts (14)	Low
Outcome: Infection-related mortality									
Downgrade -1*	No important inconsistency	No uncertainty	Imprecision	No publication bias	14/542 = 2.6%	33/480 = 6.9%	0.38 (0.21 to 0.69)	1,022 pts (10)	Low
Outcome: Febrile patients and episodes									
Downgrade -1†	Important inconsistency -1†	No uncertainty	No imprecision	No publication bias	369/708 = 52.1%	505/701 = 72.1%	0.67 (0.56 to 0.81)	1,409 pts (18)	Low
Outcome: Clinically documented infections									
Downgrade -1*	Important inconsistency -1†	No uncertainty	No imprecision	No publication bias	137/561 (24.4%)	234/558 (41.9%)	0.53 (0.36 to 0.80)	1,119 pts (14)	Low

*Some of the studies had unclear concealment of allocation, lack of blinding and unclear follow-up.

†Considerable heterogeneity: $I^2 = 86\%$.

from control to treatment would result in a confidence interval that overlaps no effect, highlighting the tenuousness of the finding and the need to downgrade for imprecision (see Table 4).³¹

Publication Bias

A systematic review with small studies, all with a positive effect and funded by industry, would raise a high degree of suspicion for the presence of publication bias that would downgrade quality. A very asymmetric funnel plot can be another indicator for publication bias. This was observed, for example, in a meta-analysis with a large number of relatively small studies (22 studies with an average of 157 patients per trial) that addressed the use of ESAs to achieve a hematologic response in hematological malignancy-related anemia ($P < .005$).³³ This asymmetric funnel plot demonstrates a number of small positive trials without the expected corresponding small negative trials and suggests that negative trials may be under-reported and thus that the apparent estimate of effect is biased upwards. However, as a word of caution, to read asymmetry in funnel plots from post hoc analysis merely as publication bias carries a considerable risk of misinterpretation.³⁴

Upgrading the Quality of Evidence

The GRADE system also recognizes rare situations, in particular for evidence from observational studies, that can upgrade the quality of the evidence. These include large and very large treatment effects and a clear dose-response relationship.

Upgrading for Large Effects

Empirical evidence³⁵ suggests that bias in observational studies tends to inflate treatment effects. However, large (RR of 0.5 or 2) or very large (RR of 0.2 or 5) treatment effects in otherwise well-done observational studies are unlikely to be explained completely by confounding (that is, imbalance of prognostic factors between intervention and control groups). As a result, the quality of evidence when effects are large can be upgraded by one or even two levels. For example, a meta-analysis from observational studies about oral anticoagulation in patients with mechanical heart valves found a relative risk reduction for thromboembolic events of more than 80%. While the observational studies are likely to overestimate the true effect, especially for beneficial effects (although less so for harm),³⁶ the study design is unlikely to explain the large effect.³⁷

Dose-Response Relationship

A clear gradient for a dose-response relationship, that is, the higher the dose, the more effective a drug or the higher the incidence of adverse events, increases our confidence that the effect is real and increases our confidence in the robustness of the effect. For example, in patients with childhood acute lymphoblastic leukemia, the risk for subsequent malignancies of the cerebral nervous system 15 years after treatment increases with the dosage of cranial radiotherapy: from 1% without radiation therapy (95% CI, 0% to 2.1%) to 1.6%

with a radiation dose of 12 Gy (95% CI, 0% to 3.4%) to 3.3% with a radiation dose of >18 Gy (95% CI, 0.9% to 5.6%).³⁸

The GRADE Evidence Profile

Guideline users can look forward to an innovation that will make it far easier for them to quickly grasp the relevant evidence supporting a treatment recommendation. GRADE has designed and tested a table format that summarizes, in a transparent fashion, the entire assessment process on a single page (see Table 4). This explicit approach enables the user to comprehend and reconstruct the process of the guideline panel and confirm—or discard—the judgments the panel has made.

Discussion

In a world of rapidly changing evidence and information overload, physicians want up-to-date guidance for the management of their patients. In an emerging world of unlimited access to healthcare information, patients want to be assured that the information they receive is rigorous and robust. In order to satisfy these needs, clinical practice guidelines must provide clear and transparent recommendations for the best care of patients, and a simple, easy-to-grasp presentation of results and recommendations for clinicians and patients alike. Building on widespread experience from the international guideline community, the GRADE framework has addressed these requirements and provides a tool for high-quality guidelines that will enable clinicians to enhance their evidence-based care.

Some people feel that guidelines should reach similar conclusions when using the same evidence base. We disagree. GRADE acknowledges the shades of gray in the evidence that often leave considerable scope for judgment and interpretation. Even more important, circumstances, values, and preferences differ across groups and jurisdictions, and the same evidence—and its interpretation—may lead to very different recommendations (for example, in wealthy and poor countries). GRADE emphasizes the need to be transparent and explicit about the judgments made in rating the quality of evidence and the values and preferences that drove the recommendations.

In some important areas of guideline development, the GRADE system pinpoints our current lack of knowledge. Guideline panels have only recently begun to include patients, their caregivers, and members of the community as legitimate members. Patient participation may be most important when treatments have severe adverse effects and the overall benefit of the therapy is often uncertain or small, as is frequently the case in hematology and oncology. Recently, authors have suggested explicit guidance regarding how to integrate patient values and preferences,³⁹ but evidence to guide the optimal approach nevertheless remains meager. Limited evidence suggests that patients and clinicians can differ widely in the way they value outcomes associated with alternative management strategies,⁴⁰ but little is understood about the circumstances in which patients and physicians are

more likely to agree than disagree. Whose values should drive recommendations: the patient, the caretaker, the community? We need to continue to explore patients and community values and preferences, and test for integrating their values and preferences in practice guidelines.³⁹

Guidelines generally address not only therapeutic but also diagnostic interventions; guidance on optimal approaches to making recommendations concerning diagnostic testing remains limited. Exploration designed to advance the current framework for evaluating diagnostic interventions is currently under way.⁴¹ The same is true for integrating issues of resource use (cost) in guidelines⁴²: recently developed guidance requires empirical testing.

While the GRADE framework has not yet solved all the issues in guideline development, the widespread international endorsement and adoption by recognized guideline organizations and medical societies such as the WHO, the American College of Physicians, the American Thoracic Society, the National Institute for Health and Clinical Excellence (NICE) UK, and organizations such as the Cochrane Collaboration or medical resources such as UpToDate, suggests that the guideline community finds the framework helpful.

Conclusion

The GRADE framework is founded in a rigorous and transparent methodology for assessing evidence, balancing benefits and harms, acknowledging values and preferences underlying specific recommendations, and integrating considerations of resource use. Transparency of each step in guideline development enables clinicians and patients to better understand and integrate the recommendations in the care of individual patients.

Acknowledgment

The authors acknowledge the members of the GRADE working group who participated over the last years in developing the framework.

References

- Schunemann HJ, Best D, Vist G, Oxman AD: Letters, numbers, symbols and words: How to communicate grades of evidence and recommendations. *CMAJ* 169:677-680, 2003
- Shaneyfelt T, Mayo-Smith M, Rothwangl J: Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA* 281:1900-1905, 1999
- Developing a Methodology for Drawing up Guidelines on Best Medical Practice (Recommendations Rec(2001)13) adopted by the Committee of Ministers of the Council of Europe. Recommendation Rec(2001). Strasbourg, France, Council of Europe Publishing, 2001
- West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al: Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No 47. Rockville, MD, Agency for Healthcare Research and Quality, 2002
- www.gradeworkinggroup.org, accessed May 2008
- Charles C, Gafni A, Whelan T, O'Brien MA: Treatment decision aids: conceptual issues and future directions. *Health Expect* 8:114-125, 2005
- Whelan T, Levine M, Willan A, Gafni A, Sanders K, Mirsky D, et al: Effect of a decision aid on knowledge and treatment decision making for breast cancer surgery: A randomized trial. *JAMA* 292:435-441, 2004
- Combination chemotherapy versus melphalan plus prednisone as treatment for multiple myeloma: An overview of 6,633 patients from 27 randomized trials. Myeloma Trialists' Collaborative Group. *J Clin Oncol* 16:3832-3842, 1998
- Alexanian R, Dimopoulos MA, Delasalle K, Barlogie B: Primary dexamethasone treatment of multiple myeloma. *Blood* 80:887-890, 1992
- Buller HR, Agnelli G, Hull RD, Hyers TM, Prins MH, Raskob GE: Antithrombotic therapy for venous thromboembolic disease: The Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 126:401-428, 2004 (suppl)
- Interferon as therapy for multiple myeloma: an individual patient data overview of 24 randomized trials and 4012 patients. *Br J Haematol* 113:1020-1034, 2001
- Ludwig H, Fritz E, Neuda J, Durie BG: Patient preferences for interferon alfa in multiple myeloma. *J Clin Oncol* 15:1672-1679, 1997
- WHO Rapid Advice Guidelines on pharmacological management of humans infected with avian influenza A (H5N1) virus. www.who.int; accessed Oct 18, 2006
- Center for Drug Evaluation and Research. Information for Healthcare Professionals. Erythropoiesis Stimulating Agents (ESA). US Food and Drug Administration. <http://www.fda.gov/cder/drug/InfoSheets/HCP/RHE200711HCP.htm>; accessed Jan 2, 2008
- Rodgers GM, et al: NCCN Practice Guidelines in Oncology: Cancer and Treatment related Anemia. Version 2.2006. NCCN National Comprehensive Cancer Network www.nccn.org/professionals/physician_gls/PDF/anemia.pdf; accessed January 2007
- Wilson J, Yao GL, Raftery J, Bohlius J, Brunskill S, Sandercock J, et al: A systematic review and economic evaluation of epoetin alpha, epoetin beta and darbepoetin alpha in anaemia associated with cancer, especially that attributable to cancer treatment. *Health Technol Assess* 11: 1-4, 2007
- Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials. *Lancet* 365:1687-1717, 2005
- Verhelst D, Rossert J, Casadevall N, Kruger A, Eckardt KU, Macdougall IC: Treatment of erythropoietin-induced pure red cell aplasia: A retrospective study. *Lancet* 363:1768-1771, 2004
- Djulfbegovic B, Wheatley K, Ross J, Clark O, Bos G, Goldschmidt H, et al: Bisphosphonates in multiple myeloma. *Cochrane Database Syst Rev* CD003188, 2002
- Clark OA, Lyman GH, Castro AA, Clark LG, Djulfbegovic B: Colony-stimulating factors for chemotherapy-induced febrile neutropenia: A meta-analysis of randomized controlled trials. *J Clin Oncol* 23:4198-4214, 2005
- Cornelissen JJ, van Putten WL, Verdonck LF, Theobald M, Jacky E, Daenen SM, et al: Results of a HOVON/SAKK donor versus no-donor analysis of myeloablative HLA-identical sibling stem cell transplantation in first remission acute myeloid leukemia in young and middle-aged adults: Benefits for whom? *Blood* 109:3658-3666, 2007
- Landolfi R, Marchioli R, Kutti J, Gisslinger H, Tognoni G, Patrono C, et al: Efficacy and safety of low-dose aspirin in polycythemia vera. *N Engl J Med* 350:114-124, 2004
- Sood AR, Burry LD, Cheng DK: Clarifying the role of rasburicase in tumor lysis syndrome. *Pharmacotherapy* 27:111-121, 2007
- Bohlius J, Reiser M, Schwarzer G, Engert A: Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. *Cochrane Database Syst Rev* CD003189, 2004
- Kunz R, Friedrich C, Wolbers M, Mann JFE: Meta-analysis: Effect of monotherapy and combination therapy with inhibitors of the renin-angiotensin system on proteinuria in renal disease. *Ann Intern Med* 148:30-48, 2008
- Parfrey PS: Inhibitors of the renin-angiotensin system: Proven benefits, unproven safety. *Ann Intern Med* 148:76-77, 2008
- Besarab A, Bolton WK, Browne JK, Egrie JC, Nissenson AR, Okamoto DM, et al: The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. *N Engl J Med* 339:584-590, 1998

28. Guyatt G, Rennie D: *Users' Guides to the Medical Literature. A Manual for Evidence-Based Clinical Practice* (ed 2). New York, NY, McGraw-Hill (in press)
29. Adjuvant chemotherapy in invasive bladder cancer: A systematic review and meta-analysis of individual patient data. Advanced Bladder Cancer (ABC) Meta-analysis Collaboration. *Eur Urol* 48:189-199, 2005
30. Wilcox RA, Djulbegovic B, Moffitt HL, Guyatt GH, Montori VM: Randomized trials in oncology stopped early for benefit. *J Clin Oncol* 26:18-19, 2008
31. Gafer-Gvili A, Fraser A, Paul M, Leibovici L: Meta-analysis: antibiotic prophylaxis reduces mortality in neutropenic patients. *Ann Intern Med* 142:979-995, 2005
32. Ross SD, Allen IE, Henry DH, Seaman C, Sercus B, Goodnough LT: Clinical benefits and risks associated with epoetin and darbepoetin in patients with chemotherapy-induced anemia: A systematic review of the literature. *Clin Ther* 28:801-831, 2006
33. Bohlius J, Wilson J, Seidenfeld J, Piper M, Schwarzer G, Sandercock J, et al: Erythropoietin or darbepoetin for patients with cancer. *Cochrane Database Syst Rev* 3:CD003407, 2006
34. Ioannidis JP, Trikalinos TA: The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *CMAJ* 176:1091-1096, 2007
35. Kunz R, Vist G, Oxman AD: Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev* 2:MR000012, 2007
36. Papanikolaou PN, Christidi GD, Ioannidis JP: Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ* 174:635-641, 2006
37. Cannegieter SC, Rosendaal FR, Briet E: Thromboembolic and bleeding complications in patients with mechanical heart valve prostheses. *Circulation* 89:635-641, 1994
38. Loning L, Zimmermann M, Reiter A, Kaatsch P, Henze G, Riehm H, et al: Secondary neoplasms subsequent to Berlin-Frankfurt-Munster therapy of acute lymphoblastic leukemia in childhood: Significantly lower risk without cranial radiotherapy. *Blood* 95:2770-2775, 2000
39. Schunemann HJ, Fretheim A, Oxman AD: Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. *Health Res Policy Syst* 4:22, 2006
40. Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, et al: Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: Observational study. *BMJ* 323:1218-1222, 2001
41. Schunemann H, Oxman A, Brozek JL, Glasziou P, Jaeschke R, Vist G, et al: GRADEing the quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* (in press)
42. Guyatt G, Oxman A, Kunz R, Jaeschke R, Helfand M, Vist G, et al: Grading recommendations: Incorporating considerations of resources use. *BMJ* (in press)