Review

# A Case for Clarity, Consistency, and Helpfulness: State-of-the-Art Clinical Practice Guidelines in Endocrinology Using the Grading of Recommendations, Assessment, Development, and Evaluation System

Brian A. Swiglo, M. H. Murad, Holger J. Schünemann, Regina Kunz, Robert A. Vigersky, Gordon H. Guyatt, and Victor M. Montori

Knowledge and Encounter Research Unit (B.A.S., M.H.M., V.M.M.), Divisions of Endocrinology, Preventive Medicine, and Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota 55905; Clinical Advances Through Research And Information Translation Research Group (H.J.S., G.H.G.), Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, Ontario L8S4L8, Canada; Department of Epidemiology (H.J.S.), Italian National Cancer Institute "Regina Elena," 00161 Rome, Italy; Basel Institute for Clinical Epidemiology (R.K.), University Hospital Basel, CH-4031 Basel, Switzerland; and Diabetes Institute (R.A.V.), Walter Reed Health Care System, Washington, D.C. 20307

**Context:** The Endocrine Society, and a growing number of other organizations, have adopted the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) system to develop clinical practice guidelines and grade the strength of recommendations and the quality of the evidence. Despite the use of GRADE in several of The Endocrine Society's clinical practice guidelines, endocrinologists have not had access to a context-specific discussion of this system and its merits.

**Evidence Acquisition:** The authors are involved in the development of the GRADE standard and its application to The Endocrine Society clinical practice guidelines. Examples were extracted from these guidelines to illustrate how this grading system enhances the quality of practice guidelines.

**Evidence Synthesis:** We summarized and described the components of the GRADE system, and discussed the features of GRADE that help bring clarity and consistency to guideline documents, making them more helpful to practicing clinicians and their patients with endocrine disorders.

**Conclusions:** GRADE describes the quality of the evidence using four levels: very low, low, moderate, and high quality. Recommendations can be either strong ("we recommend") or weak ("we suggest"), and this strength reflects the confidence that guideline panel members have that patients who receive recommended care will be better off. The separation of the quality of the evidence from the strength of the recommendation recognizes the role that values and preferences, as well as clinical and social circumstances, play in formulating practice recommendations. **(*J Clin Endocrinol Metab* 93: 666–673, 2008)**

Professional organizations, such as The Endocrine Society and its sister societies, have set out to develop clinical practice guidelines to provide helpful recommendations to practicing clinicians, to improve quality of care, and to enhance patient outcomes. By producing guidelines, these organizations seek to assert their academic and practice leadership in areas of primary concern. Given the policy and legal implications of guidelines, state-of-the-art guideline developers follow rigorous and transparent procedures for formulating recommendations for or against a particular diagnostic or therapeutic intervention.

0021-972X/08/$15.00/0

Printed in U.S.A.



doi: 10.1210/jc.2007-1907 Received August 24, 2007. Accepted December 21, 2007.

First Published Online January 2, 2008

Abbreviations: GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; RCT, randomized controlled trial.

Guidelines are strengthened further if they involve panel members without substantial conflicts of interest (*i.e.* members who do not expect to benefit directly or indirectly, now or in the future, personally or financially, from making a particular recommendation) and conduct their proceedings without for-profit support. Key to their success is the expectation that clinicians will deliver better care for their patients if they follow guideline recommendations. Thus, clinicians need to find the recommendations both clear and helpful.

In this article we will discuss the processes involved in developing helpful and rigorous clinical practice guidelines in a manner congruent with the approach The Endocrine Society has adopted. We anticipate that this will assist endocrinologists and other parties who are interested in critically appraising, implementing, and enhancing The Endocrine Society's clinical practice guidelines.

## Developing Rigorous and Helpful Clinical Practice Guidelines

Evidence-based medicine recognizes two principles (1). The first is that there is a hierarchy of evidence such that one is more confident about decisions based on evidence that offers greater protection against bias and random error. The second principle is that evidence alone is never sufficient to make clinical decisions. In fact, evidence-based medicine stipulates that optimal treatment decisions require integration of clinical knowledge and research evidence with patient circumstances, including their values and preferences. The rigorous application of these principles to the development of clinical practice guidelines is a relatively recent development.

Therefore, evidence-based guidelines are most helpful when they provide recommendations that are clear, based on the best available research evidence, and transparent in terms of reporting the quality of the evidence and the basis for determining the strength of the recommendations. Often this includes explicitly describing the pertinent values and preferences the guideline authors bring to bear in developing the recommendations.

For over a decade, most guideline groups have recognized that developing a summary categorization of the strength of the recommendations and the quality of the evidence supporting them, processes sometimes called grading (of the recommendation strength) and rating (of the evidence quality), helps clinicians understand a practice guideline's summary message. Multiple systems in use produce different grading and rating categories, and rely on different letters, numbers, symbols, and terms (2). This can cause confusion while clarity is needed.

To address this concern, the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) working group, comprised of expert methodologists and guideline developers from a variety of health care organizations, set out to: 1) evaluate these different systems, 2) develop one recommended grading system, and 3) disseminate this system throughout medical communities and their literature. The challenge was great because many systems were already in place, all systems have limitations, and many organizations have spent significant resources on developing their rating system (3). GRADE's design criteria included simplicity and applicability to a wide variety of clinical recommendations that encompass the full spectrum of patient management decisions. The GRADE working group first published their findings in 2004 (4).

Since that time, numerous organizations have adopted GRADE as their guideline grading system. These organizations include The Endocrine Society, World Health Organization, American College of Chest Physicians, UpToDate, American College of Physicians, American Thoracic Society, The Cochrane Collaboration, European Respiratory Society, Agency for Healthcare Research and Quality, and Society of Critical Care Medicine (a complete list is available on the GRADE working group web site) (5). An emerging consensus seems to be forming around the adoption of GRADE. This would be a welcome progression because such widespread adoption will help maintain clarity and consistency in guidelines across medical disciplines.

The Endocrine Society appraised the merits of the GRADE system and decided in late 2004 to adopt it as the basis for its clinical practice guidelines. The Endocrine Society was the first North American organization to adopt GRADE and use it in its Clinical Practice Guidelines program. Guidelines on the use of testosterone in men (6), on the treatment and prevention of pediatric obesity, and on the diagnosis and treatment of hirsutism are examples of the application of the GRADE system to The Endocrine Society guidelines. However, endocrinologists have not had access to a context-specific discussion of this system as it relates to guidelines in endocrinology. In the following sections, we will use endocrinology examples to illustrate how this grading system helps improve the rigor and usefulness of clinical practice guidelines.

## The GRADE System

The GRADE system classifies recommendations into one of two grades (strong or weak) and the quality of the evidence into one of four categories (high, moderate, low, or very low). This offers a simple and practical, yet methodologically rigorous, grading system for The Endocrine Society's Clinical Practice Guidelines program.

To enhance further the interpretation and clarity of the recommendations, guideline developers use the terms "we recommend" to denote strong recommendations, whereas weak recommendations use the less definitive wording "we suggest." Furthermore, a strong recommendation receives a grade 1 classification, and a weak recommendation receives a grade 2 classification. The symbols chosen for the four levels of quality of evidence are: ⊕◯◯◯ (very low); ⊕⊕◯◯ (low); ⊕⊕⊕◯ (moderate); and ⊕⊕⊕⊕ (high) quality. Table 1 provides an overview of the GRADE system and a closer look at the components of each of its recommendation categories.

## Strength of Recommendations

The strength of a recommendation reflects the degree of confidence that the desirable effects of a recommendation outweigh

**TABLE 1.** GRADE recommendations–a closer look

| Rating of evidence quality | Clarity of risk/benefit | Description of supporting evidence | Implications |
|---|---|---|---|
| Strong recommendations | | | |
| High quality evidence | Benefits clearly outweigh harms and burdens, or vice versa | Consistent evidence from well-performed RCTs or exceptionally strong evidence from unbiased observational studies[a] | Recommendation can apply to most patients in most circumstances. Further research is very unlikely to change our confidence in the estimate of effect. |
| Moderate quality evidence | Benefits clearly outweigh harms and burdens, or vice versa | Evidence from RCTs with important limitations (inconsistent results, methodological flaws, indirect or imprecise evidence), or unusually strong evidence from unbiased observational studies | Recommendation can apply to most patients in most circumstances. Further research (if performed) is likely to have an impact on our confidence in the estimate of effect and may change the estimate. |
| Low quality evidence | Benefits clearly outweigh harms and burdens, or vice versa | Evidence for at least one critical outcome from observational studies, from RCTs with serious flaws, or indirect evidence | Recommendation may change when higher quality evidence becomes available. 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 evidence (very rarely applicable) | Benefits clearly outweigh harms and burdens, or vice versa | Evidence for at least one of the critical outcomes from unsystematic clinical observations or very indirect evidence | Recommendation may change when higher quality evidence becomes available; any estimate of effect, for at least one critical outcome, is very uncertain. |
| Weak recommendations | | | |
| High quality evidence | Benefits closely balanced with harms and burdens | Consistent evidence from well-performed RCTs or exceptionally strong evidence from unbiased observational studies | The best action may differ depending on circumstances or patient or societal values. Further research is very unlikely to change our confidence in the estimate of effect. |
| Moderate quality evidence | Benefits closely balanced with harms and burdens | Evidence from RCTs with important limitations (inconsistent results, methodological flaws, indirect or imprecise evidence), or unusually strong evidence from unbiased observational studies | Alternative approaches likely to be better for some patients under some circumstances. Further research (if performed) is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. |
| Low quality evidence | Uncertainty in the estimates of benefits, harms, and burdens; benefits may be closely balanced with harms and burdens | Evidence for at least one critical outcome from observational studies, from RCTs with serious flaws, or indirect evidence | Other alternatives may be equally reasonable. 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 evidence | Major uncertainty in the estimates of benefits, harms, and burdens; benefits may or may not be balanced with harms and burdens | Evidence for at least one critical outcome from unsystematic clinical observations or very indirect evidence | Other alternatives may be equally reasonable. Any estimate of effect, for at least one critical outcome, is very uncertain. |

Modified from Schunemann *et al.* (22).

[a] Exceptionally strong evidence from unbiased observational studies includes: 1) evidence from studies that yield estimates of the treatment effect that are large and consistent; 2) evidence in which all potential biases may be working to underestimate an apparent treatment effect, and therefore, the actual treatment effect is likely to be larger than that suggested by the study data; and 3) evidence in which a dose-response gradient exists.

the undesirable effects. Desirable effects can include beneficial health outcomes, less burden, and cost savings. Undesirable effects can include harms, more burden, and expenses. Burdens are the demands of adhering to a recommendation that patients or caregivers (*e.g.* their family) may dislike, such as having to take medication or the inconvenience of going to the doctor's office. Although the degree of confidence is a continuum, the GRADE approach classifies recommendations for or against treatments into two grades, strong and weak.

If guideline developers are confident that the desirable effects of adherence to a recommendation outweigh the undesirable effects, they will make a strong recommendation within the con-

text of a described intervention. Typically, this requires high or moderate quality evidence on patient important outcomes. Exceptionally, panels can make strong recommendations based on low to very low quality evidence. This may occur when the values and preferences guideline developers bring to bear are such that when considering even low quality evidence, they are confident that the benefits of an intervention outweigh the undesirable outcomes (or vice versa). In these cases the panel can make a strong recommendation for (or against) the intervention.

For example, consider the decision to administer aspirin or acetaminophen to children with chicken pox. Observational studies have noted an association between aspirin administration and Reye syndrome. Because aspirin and acetaminophen are, in this context, similar in their analgesic and antipyretic effects, guideline developers may make a strong recommendation for acetaminophen despite the low quality evidence suggesting harm from aspirin because they place a very high value on avoiding potential life-threatening adverse effects.

A weak recommendation is one for which a guideline panel concludes that the desirable effects of adherence to a recommendation probably outweigh the undesirable effects, but the panel is not confident. Thus, if guideline developers believe that benefits and downsides are finely balanced, or appreciable uncertainty exists about this balance, they offer a weak recommendation. Thus, low or very low quality evidence usually leads to weak recommendations because of uncertainty about the balance between risks and benefits. Guideline panels may offer weak recommendations even when high quality evidence is available when that evidence clearly demonstrates that the benefits and risks are closely balanced. For example, a guideline panel may weakly recommend bisphosphonates in relatively low-risk patients with osteopenia, in whom the burden and costs of monitoring and treatment may or may not be worth the potential reduction in the risk of fragility fractures documented in randomized trials.

Table 2 summarizes the factors that influence the strength of a recommendation, factors that broadly correspond to: 1) certainty about the balance between benefits *vs*. burdens and harms, 2) resource use, and 3) variation in values and preferences. Consideration of this latter issue is key. Guideline panels will typically, either explicitly or implicitly, use their own preferences as imperfect proxies of patient values. Alternatively, they could consider the range of patients to whom the recommendation applies, and their range of values and preferences. Ideally, they will find a way to ensure that the recommendation is consistent with the values and preferences of most patients. How to achieve this goal remains a challenge; one approach includes involving relevant patients as panel members or involving patient groups able to minimize influences that could bias their judgments in the assessment of values and preferences.

There are practical implications relating the strength of recommendation for or against a therapy with patient values and

**TABLE 2.** Factors in deciding on a strong or weak recommendation

| Issue | Endocrinology example |
|---|---|
| Methodological quality of the evidence supporting estimates of desirable and undesirable outcomes | Many high quality RCTs consistently show a large reduction in cardiovascular risk among patients with diabetes treated with statins; only unsystematic clinical observations support the type of glucocorticoid preparation used in patients with adrenal insufficiency. |
| Importance of the outcome that treatment prevents | Use of testosterone therapy in young patients with severe hypogonadism to improve their quality of life *vs*. use of testosterone in asymptomatic, healthy elderly men with low testosterone levels to normalize these levels. |
| Magnitude of treatment effect | When compared with conventional control, tight blood pressure control reduces the risk of diabetes-related complications to a greater extent (RRR 24%) than tight glycemic control (RRR 12%) in patients with type 2 diabetes. |
| Precision of the estimate of treatment effect | The relative risk associated with calcitonin for the prevention of vertebral fractures has a wider confidence interval than the relative risk associated with bisphosphonates. |
| Risks associated with therapy | Metformin reduces hemoglobin A1c with much lower risk of hypoglycemia than sulfonylureas. |
| Burdens of therapy | In patients with newly diagnosed type 2 diabetes, insulin therapy is associated with a higher burden than taking metformin; optimal insulin use requires regular glucose self-monitoring and deliberate food intake. |
| Risk for target event | Elderly women with a personal history of a fragility fracture have a much higher risk of another fragility fracture than otherwise healthy elderly women with low bone density. |
| Resource use | Teriparatide is a more expensive treatment, and, thus, implies greater resource use, for secondary prevention of osteoporotic fractures than oral bisphosphonates. |
| Varying values | Most young, healthy persons with hyperparathyroidism may put a high value on avoiding prolonged medical monitoring and risk of complications (kidney stones, bone fractures) and a low value on surgical risk, and, thus, may prefer to undergo parathyroidectomy; many elderly and frail patients may put a high value on avoiding surgical risk and a low value in avoiding periodic monitoring, and, thus, may refuse surgery. |

From Guyatt *et al*. (3). RRR, Relative risk reduction.

preferences. For instance, a strong recommendation implies that virtually all patients, across the range of individual values found in the population, will make the same treatment decision. Strong recommendations allow clinicians to offer treatment with confidence, commonly with limited to no consideration of alternative options. Weak recommendations imply that different patients, in different clinical contexts, with different values and preferences, will likely make different choices. In the face of weak recommendations, clinicians will need to be more deliberate and judicious in explicitly incorporating evidence regarding the magnitude of benefits and risks along with patient circumstances, values, and preferences to make the best decision. In other words, with weak recommendations, the clinician will need to have a more detailed and deliberate discussion with the patient, reviewing several reasonable options. This is particularly important when clinicians and patients find their own values and preferences at odds with those the guideline panel considered in making its recommendations.

We do not know how individual clinicians can best achieve the goal of incorporating patient values and preferences in following a weak recommendation, but some promising approaches exist. For example, some clinicians are using decision aids. Decision aids are tools that help clinicians communicate to patients the relevant evidence about the available options and their relative merits in a quantitative form. Examples of these tools can be found elsewhere (for examples, see http://kerunit.e-bm.org). Randomized trials have shown that these tools can improve the quality of decision making in many clinical settings (7). Conversely, for strong recommendations, a decision aid could be an inefficient use of time and other resources; although it is plausible that having the patient participate in making treatment choices may enhance adherence to therapy (8).

## Quality of the Evidence

To determine the strength of the recommendations, the GRADE system explicitly considers the quality of the best available evidence identified through a comprehensive review of the literature. Study design and conduct are important determinants of quality. Randomized controlled trials (RCTs) allow decision makers to draw causal inferences linking interventions and outcomes with protection against bias. Therefore, RCTs begin with a "high" quality rating.

Because of possible limitations that fall into five categories (Table 3), even RCTs may not provide high quality evidence. First, there may be serious limitations in the design and conduct of RCTs (including lack of concealment and blinding, and large loss to follow-up), and these limitations would lead to a reduction in the quality of the evidence base (weakening the inference decision makers can draw from these data) and in turn a reduction in the quality level. For example, to inform guideline developers about the efficacy of physical activity on pediatric obesity, the authors reviewed the results of a metaanalysis of 20 relevant RCTs (9). These trials had no reported allocation concealment or blinding and had significant loss to follow-up (29% of studies reported greater than a 20% loss). Therefore, the guideline panel downgraded the quality of the evidence.

**TABLE 3.** Factors in deciding on confidence in estimates of benefits, risks, burdens, and costs

| |
|---|
| Factors that may decrease the quality of evidence |
| Poor quality of planning and implementation of the available RCTs suggesting high likelihood of bias |
| Inconsistency of results |
| Indirectness of evidence |
| Lack of precision; sparse evidence |
| Reporting bias (including publication bias) |
| 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 |

From Guyatt *et al.* (3).

Second, if the results of trials are highly variable, we will have less confidence in the estimates of efficacy, and the evidence will have lower quality. For instance, RCTs of testosterone use in adult men reveal an inconsistent effect on lumbar spine bone mineral density and on libido (in trials that enrolled men with low testosterone levels) (10). These findings lower the quality of the evidence. However, in the first case, a planned subgroup analysis revealed a significant and large interaction between the route of administration and the treatment effect, explaining the inconsistency, and increasing the confidence of the guideline developers about the effect of intramuscular testosterone on lumbar spine bone mineral density (Figs. 1 and 2) (10). Thus, guideline developers did not need to downgrade the quality rating for this evidence because of inconsistency. In contrast, developers downgraded the evidence linking testosterone use and libido because of unexplained and very large inconsistency (11).

Third, a reduction in quality will occur when evidence supporting a recommendation is indirect. Indirectness may occur if: the patients enrolled in relevant trials differ in important ways from those under consideration by the guideline panel; the intervention or the comparator intervention tested in the trials differ in important ways (nature, dosing, duration) from those under consideration; or the outcomes differ (typically investigators will have measured effects on a substitute or surrogate outcome, rather than the patient-important outcome in which the guideline panel is primarily interested).

For example, when considering the use of testosterone gel to prevent fragility fractures in elderly hypogonadal men, evidence from trials enrolling younger men show that intramuscular testosterone can increase bone mineral density (12, 13). Here, the evidence informs the efficacy of a different testosterone formulation on a different patient group on a surrogate outcome of no importance, in and of itself, to patients (bone density rather than fracture risk); no high-quality trials have answered the question of direct relevance to the guideline developer. If a recommendation was made specific to the use of testosterone gel to prevent fractures in elderly men, the quality of the evidence would be downgraded based on indirectness with respect to the population, intervention, and outcome. Furthermore, the guideline panel interested in making recommendations about the use of

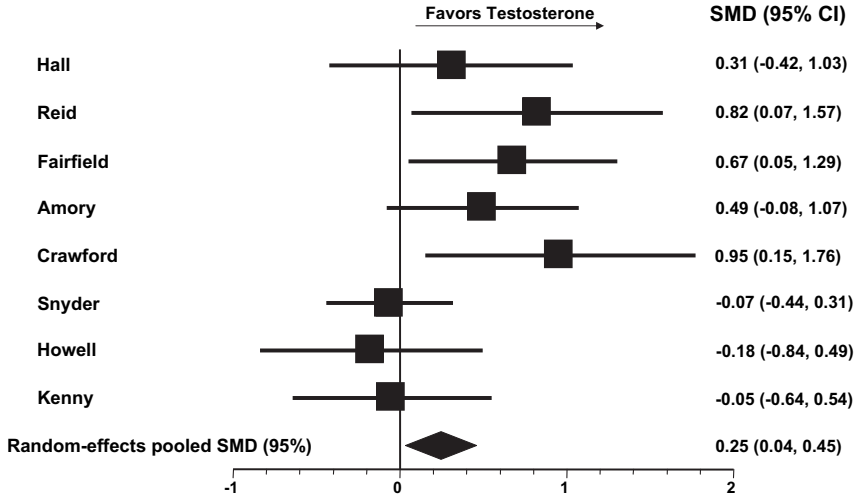


**FIG. 1.** Inconsistent results. This displays random-effects metaanalysis results of eight trials of testosterone on lumbar spine bone mineral density. $I^2$ (a statistic that reflects the proportion of variation between studies that is not due to chance, *i.e.* inconsistency) is 46%, which identifies substantial inconsistency. *Vertical line* indicates no treatment effect; *squares* and *horizontal lines* indicate point estimates and associated 95% confidence intervals (CIs) for each study. *Diamonds* indicate the random-effects pooled standard mean difference (SMD) with the width representing its confidence interval (10).

testosterone for osteoporosis will have to rely only on indirect comparisons (*i.e.* trials of each agent against placebo but no head-to-head trials) when considering the relative merits of testosterone *vs.* bisphosphonates, for instance.

Fourth, guideline developers should downgrade evidence when few studies, involving few participants and, most importantly, documenting few outcomes, inform the tradeoffs of risks and benefits. As an example, a metaanalysis of the results of trials evaluating the effects of testosterone on cardiovascular outcomes suggests that testosterone does not have an effect on cardiovascular events. However, this result is based on only six trials, a total of 308 participants, and only 21 outcomes. Considering the confidence interval width, the pooled data are consistent with both a 1-fold decrease and a 4-fold increase in the odds of cardiac events in patients treated with testosterone (14). This evidence carries great uncertainty, lowering the confidence that the estimates are accurate.

Finally, guideline developers should have limited confidence when reporting bias might have affected the underlying evidence. Publication bias, one form of reporting bias, occurs because trials that show no significant effect are less likely to be published, and outcome reporting bias occurs when researchers selectively report their findings depending on their significance. Clinical trial registries may help reduce publication bias (15). Chan *et al.* (16) found that reporting of trial outcomes is frequently incomplete, biased, and inconsistent with the original trial protocols. Prospective public registration of trial protocols could help diminish this concern. Box 1 describes an example of reporting bias. Publication bias is more likely to take place in fields in which small trials are the norm (*e.g.* many endocrinopathies) because large trials are less likely to remain unpublished. Although difficult to ascertain, reporting bias is prevalent, particularly when key patient-important outcomes are only reported in a few studies.

In contrast to RCTs, observational studies start with a "low" (*i.e.* case-control studies, and cohort studies) or "very low" (*i.e.* unsystematic clinical observations, case reports and series) quality level but may be upgraded in certain situations, *e.g.* when the magnitude of the treatment effect is very large (*e.g.* use of insulin to prevent morbidity and mortality in patients with type 1 diabetes presenting in diabetic ketoacidosis; use of glucocorticoids to prevent adrenal crisis in patients with Addison's disease). Thus, it is very important in guidelines to specify clearly the alternatives considered. Although high quality evidence, as we have seen, supports the use of glucocorticoids to prevent adrenal crises in patients with Addison's disease, low quality evidence supports the choice of a specific glucocorticoid replacement regimen out of several in common use.

In addition, the quality level can increase when all plausible confounders would reduce the magnitude of the treatment effect, yet the effect remains sizeable. For example, a systematic review showed higher mortality in for-profit hospitals when compared with not-for-profit hospitals (17). This result occurred despite the fact that for-profit hospitals usually have additional resources available and generally admit healthier patients, factors that should work in their favor. Considering these confounders would increase the magnitude of benefit of not-for-profit hospitals (3). Table 3 summarizes factors that influence the quality of evidence.

## Values and Preferences

As mentioned previously, values and preferences are essential to guidelines. The GRADE system offers insights into the role of values and preferences when it disentangles the strength of recommendations from the quality of the evidence, and encourages statements about the underlying values and preferences relevant to the recommendations.

Consider the interpretation of guidelines in the case of an individual patient. A guideline may weakly recommend (a "suggestion," using the terminology of The Endocrine Society Clinical Practice Guidelines) that patients receive treatment with a medication based on low quality evidence because there is uncertainty about the tradeoffs between potential desirable and undesirable effects. An individual patient may place a high value on potential resolution of their symptoms and a low value on avoiding possible side effects, costs, and follow-up visits and tests while taking the medication. Such a patient may prefer to take this medication, in keeping with the suggestion. Another patient in similar circumstances may have different values, placing a higher value on avoiding potential adverse effects, costs, and burdens of medical treatment.

For example, when making a decision on treatment options for the prevention of osteoporotic fractures, some experts may formulate recommendations in favor of treatment with teripa-

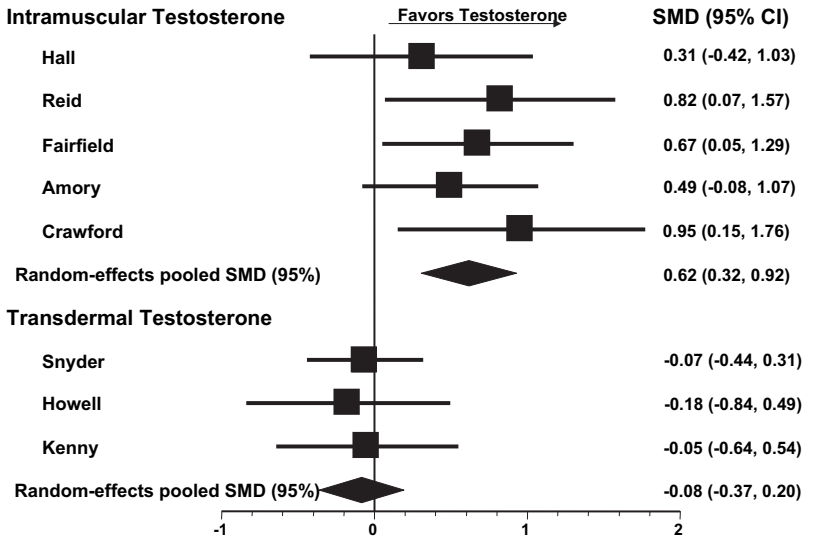


**FIG. 2.** Consistent results after subgroup analysis. This displays random-effects metaanalysis results of the same eight trials of testosterone on lumbar spine bone mineral density that are displayed in Fig. 1, but after subgroup analysis by administration route. There was no measurable inconsistency ($I^2 = 0\%$) for both subgroups. *Vertical line* indicates no treatment effect; *squares* and *horizontal lines* indicate point estimates and associated 95% confidence intervals (CIs) for each study. *Diamonds* indicate the random-effects pooled standard mean difference (SMD) with the width representing its confidence interval (10).

ratide for women at high fracture risk. One woman may share values and preferences in keeping with this recommendation, whereas another woman, in the same situation, may find the route of administration (injection) or the cost of teriparatide unacceptable and would thus prefer not to take the medication. The use of the GRADE system, with its transparency, offers patients and clinicians the opportunity to consider and make different clinical decisions, including decisions to not use an intervention that is weakly recommended (or to use one that the guideline weakly recommends against).

The appendix (published as supplemental data on The Endocrine Society's Journals Online web site at http://jcem.endojournals.org) offers illustrations from The Endocrine Society Clinical Practice Guidelines to highlight the issues presented here.

Box 1. An example of reporting bias

A systematic review of the effects of testosterone on erection satisfaction and function in patients with low testosterone offers an example of reporting bias. In this review the authors found one large trial that specifically addressed this issue in addition to three smaller trials (11). However, the large trial's results on the outcome of interest were reported only as "not significant" in the published paper; the actual data were not reported and, therefore, could not be used in a metaanalysis. Using the data from the three other trials, there was a large treatment effect noted with testosterone therapy (difference between arms of 1.3 SD values, 95% confidence interval 0.2 to 2.3). However, after obtaining the complete data on the larger trial, the new pooled treatment effect was smaller in magnitude, much less precise, and no longer significant (0.8 SD values, 95% confidence interval −0.05 to 1.63), an example of reporting bias (23).

## Future Directions

GRADE does not answer all questions related to rigorous guideline development, but many areas, such as diagnostic recommendations and consideration of resource allocation, are in active development. We anticipate updating the endocrine readership when further guidance becomes available.

In regards to considering resource allocation in guidelines, there are challenges concerning the clarity, conflicts, validity, and applicability of the evidence (*e.g.* cost-effectiveness analyses), challenges in the interpretation and use of economic analyses to formulate guidelines (without the guidance of a health economist), and the impact of such analyses when guidelines are intended for broad, or even international, audiences. The American College of Chest Physicians has suggested an approach to this problem that is consistent with GRADE (18). The GRADE working group is preparing documents and a conference that will provide additional guidance on this topic.

There is also uncertainty as to the ideal composition of the guideline panel. Some favor broad representation, expanding from the usual set of clinical experts to include patients and health officials. However, how to select patients for participation in guidelines (*e.g.* highly educated patients are likely to participate actively, but they may not share values with many other patients), how to engage them into the process, and how to acknowledge their contribution is the subject of evolving science (19–21). The promise of being able to incorporate values and preferences in guideline development through direct patient consultation seems a fascinating prospect.

## Conclusions

Guideline development processes that are adherent to the principles of evidence-based medicine, such as the GRADE system, offer clarity, transparency, consistency, and helpfulness for academic and professional organizations seeking to provide their clinicians with practice recommendations. Further experience in the use of GRADE in endocrine guidelines and familiarity of the users with this system could enhance evidence-based endocrine practice.

## Acknowledgments

We thank the leaders and members of The Endocrine Society Task Forces who have pioneered the use of Grading of Recommendations, Assess-

ment, Development, and Evaluation, and other efforts to formulate recommendations in endocrine practice.

Address all correspondence and requests for reprints to: Victor M. Montori, M.D., M.Sc., Mayo Clinic, W18A, 200 First Street SW, Rochester, Minnesota 55905. E-mail: montori.victor@mayo.edu.

Disclosure Statement: V.M.M. receives funding from The Endocrine Society to conduct systematic reviews and metaanalyses in support of clinical practice guidelines. R.A.V. chairs the Clinical Guidelines Subcommittee of The Endocrine Society. H.J.S., R.K., G.H.G., and V.M.M. are members of the Grading of Recommendations, Assessment, Development, and Evaluation Working Group. H.J.S. is funded by a European Commission Grant (The human factor, mobility and Marie Curie Actions. Scientist Reintegration Grant IGR 42192–"GRADE"). Otherwise, the authors have nothing to disclose.

## References

1. **Guyatt GH, Haynes B, Jaeschke R, Cook D, Greenhalgh T, Meade M, Green L, Naylor C, Wilson M, McAlister FA, Richardson W, Montori V, Bucher H** 2002 Introduction: the philosophy of evidence-based medicine. In: Guyatt GH, Rennie D, eds. Users' guides to the medical literature: a manual of evidence-based clinical practice. Chicago: AMA Press; 121–140
2. **Schunemann HJ, Best D, Vist G, Oxman AD** 2003 Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ [Erratum (2004) 170:1082] 169:677–680
3. **Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H** 2006 An emerging consensus on grading recommendations? ACP J Club 144:A8–A9
4. **Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams Jr JW, Zaza S** 2004 Grading quality of evidence and strength of recommendations. BMJ 328:1490
5. **GRADE working group** 2007 Organizations. Available at: http://www.gradeworkinggroup.org/society/index.htm. Accessed May 8, 2007
6. **Bhasin S, Cunningham GR, Hayes FJ, Matsumoto AM, Snyder PJ, Swerdloff RS, Montori VM** 2006 Testosterone therapy in adult men with androgen deficiency syndromes: an endocrine society clinical practice guideline. J Clin Endocrinol Metab 91:1995–2010
7. **O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, Tait V, Tetroe J, Fiset V, Barry M, Jones J** 2003 Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev (2):CD001431
8. **Weymiller AJ, Montori VM, Jones LA, Gafni A, Guyatt GH, Bryant SC, Christianson TJ, Mullan RJ, Smith SA** 2007 Helping patients with type 2 diabetes mellitus make treatment decisions: statin choice randomized trial. Arch Intern Med 167:1076–1082
9. **McGovern L, Johnson JN, Paulo R, Hettinger A, Singhal V, Kamath C, Erwin PJ, Montori VM,** Treatment of pediatric obesity. A systematic review and metaanalysis of randomized trials. J Clin Endocrinol Metab, in press
10. **Tracz MJ, Sideras K, Bolona ER, Haddad RM, Kennedy CC, Uraga MV, Caples SM, Erwin PJ, Montori VM** 2006 Testosterone use in men and its effects on bone health. A systematic review and meta-analysis of randomized placebo-controlled trials. J Clin Endocrinol Metab 91:2011–2016
11. **Bolona ER, Uraga MV, Haddad RM, Tracz MJ, Sideras K, Kennedy CC, Caples SM, Erwin PJ, Montori VM** 2007 Testosterone use in men with sexual dysfunction: a systematic review and meta-analysis of randomized placebo-controlled trials. Mayo Clin Proc 82:20–28
12. **Behre HM, Kliesch S, Leifke E, Link TM, Nieschlag E** 1997 Long-term effect of testosterone therapy on bone mineral density in hypogonadal men. J Clin Endocrinol Metab 82:2386–2390
13. **Katznelson L, Finkelstein JS, Schoenfeld DA, Rosenthal DI, Anderson EJ, Klibanski A** 1996 Increase in bone density and lean body mass during testosterone administration in men with acquired hypogonadism. J Clin Endocrinol Metabol 81:4358–4365
14. **Haddad RM, Kennedy CC, Caples SM, Tracz MJ, Bolona ER, Sideras K, Uraga MV, Erwin PJ, Montori VM** 2007 Testosterone and cardiovascular risk in men: a systematic review and meta-analysis of randomized placebo-controlled trials. Mayo Clin Proc 82:29–39
15. **Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, Godlee F, Haug C, Hebert PC, Kotzin S, Marusic A, Sahni P, Schroeder TV, Sox HC, Van der Weyden MB, Verheugt FW** 2007 Clinical trial registration–looking back and moving ahead. N Engl J Med 356:2734–2736
16. **Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG** 2004 Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 291:2457–2465
17. **Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schunemann HJ, Haines T, Lavis JN, Grant BJ, Haslam DR, Bhandari M, Sullivan T, Cook DJ, Walter SD, Meade M, Khan H, Bhatnagar N, Guyatt GH** 2002 A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ 166:1399–1406
18. **Guyatt G, Baumann M, Pauker S, Halperin J, Maurer J, Owens DK, Tosteson AN, Carlin B, Gutterman D, Prins M, Lewis SZ, Schunemann H** 2006 Addressing resource allocation issues in recommendations from clinical practice guideline panels: suggestions from an American College of Chest Physicians task force. Chest 129:182–187
19. **Fretheim A, Schunemann HJ, Oxman AD** 2006 Improving the use of research evidence in guideline development: 5. Group processes. Health Res Policy Syst 4:17
20. **Fretheim A, Schunemann HJ, Oxman AD** 2006 Improving the use of research evidence in guideline development: 3. Group composition and consultation process. Health Res Policy Syst 4:15
21. **Schunemann HJ, Fretheim A, Oxman AD** 2006 Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. Health Res Policy Syst 4:22
22. **Schunemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan KL, Krishnan JA, Manthous CA, Maurer JR, McNicholas WT, Oxman AD, Rubenfeld G, Turino GM, Guyatt G** 2006 An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. Am J Respir Crit Care Med 174:605–614
23. **Sinha MK, Montori VM** 2006 Reporting bias and other biases affecting systematic reviews and meta-analyses: a methodological commentary. Expert Rev Pharmacoeconomics Outcomes Res 6:603–611