World J Urol (2011) 29:311–317
DOI 10.1007/s00345-011-0667-2

TOPIC PAPER

Rating the quality of evidence and the strength of recommendations using GRADE

Steven E. Canfield · Philipp Dahm

Received: 9 November 2010 / Accepted: 22 February 2011 / Published online: 9 March 2011

Abstract
Objectives Urologists can benefit from a standardized system for guideline development and presentation. This article introduces the GRADE system and explains how it may be useful for Urologic physicians, in their practice and in their healthcare systems.
Methods The GRADE system is reviewed. Specific aspects of how GRADE rates the quality of the evidence and the strength of recommendations are explored.
Results GRADE can provide explicit and structured guidance, which separates the quality of evidence from the strength of recommendations. This information can be used by consumers of guidelines, including patients, physicians, and policy makers.
Conclusions Urologists can benefit from a more transparent and rigorous framework when formulating recommendations. GRADE is an emergent proposal with broader implications for healthcare policy as well.

Keywords Guidelines · Levels of evidence · Urology

Introduction

Urologic guidelines can be difficult to compare due to wide variations in how they rate the quality of evidence and present their recommendations. Often, the authors cannot offer an actual recommendation due to perceived gaps in the evidence, or when a recommendation is offered, it may be unclear what the ranking of the recommendation implies. Unfortunately, such guidelines may not provide much actual guidance and may introduce confusion. Grading of Recommendations Assessment, Development, and Evaluation (GRADE) is a system which provides clear and concise information on both the quality of the evidence and the strength of the recommendation. The system can be used when developing systematic reviews and when formulating recommendations in the context of guidelines. Information on patient important outcomes is presented in a systematic and explicit fashion which can be used by physicians, patients, and policy makers.

S. E. Canfield (✉)
Division of Urology, UT Medical School at Houston,
6431 Fannin St. MSB 6.018, Houston, TX 77030, USA
e-mail: steven.canfield@uth.tmc.edu

P. Dahm
Department of Urology, College of Medicine,
University of Florida, Gainesville, FL, USA

Why is a better system needed?

The purpose of a guideline is to provide a summary of best-known practices for a given topic, which can aid practitioners, may lead to improved outcomes for patients, and may inform healthcare policy. Currently, there is a confusing array of urological guideline formats which can be demonstrated by a brief assessment of examples from three major organizations (Table 1). Each of these organizations is internationally respected and performs high-quality evidence-based evaluations for their guideline recommendations. The Scottish Intercollegiate Guideline Network (SIGN) denotes the level of evidence as A, B, or C and provides a recommendation [1]. As an example, consider recommendations for improving urinary control after radical prostatectomy. The SIGN guideline recommends that pelvic floor muscle exercise should be considered, grade B [2]. The National Comprehensive Cancer Network (NCCN) develops algorithms which attempt to identify logical disease

Table 1 Examples of current guideline rating and grading formats

SIGN	NCCN	EAU[a]
A—randomized controlled trials (RCTs) or systematic reviews of RCTs	1—high-quality evidence along with uniform level of consensus	A—clinical studies of good quality and consistency addressing the specific recommendations and including at least one randomized controlled trial
B—non-randomized trials and other observational studies	2a and 2b—lower-quality evidence with uniform or non-uniform consensus	B—well-conducted clinical studies, but without randomized clinical trials
C—expert opinion	3—any quality evidence but with major disagreement among panelists	C—recommendations made despite the absence of directly applicable clinical studies of good quality

[a] EAU levels of evidence applied as well based on Oxford centre for evidence-based medicine levels of evidence

pathways and guideline elements (recommendations) [3]. The NCCN suggests that preserving urethral length and avoiding damage to the external sphincter can "reduce" incontinence, and preserving the bladder neck may "decrease the risk" of incontinence, category 2A [4]. The European Association of Urology (EAU) presents both levels of evidence and grades for recommendations derived from the Agency for Healthcare Research and Quality (AHRQ) [5]. They recommend that "some preoperative or immediate postoperative instructions in pelvic muscle training for men undergoing radical prostatectomy may be helpful", grade B [6].

These three examples illustrate the diverse approach undertaken by just a few of the many important organizations which seek to provide practical guidance to urologists. Yet an attempt to compare the information provided by these different groups would prove difficult. The methodology used is too different, and there is no crosswalk from guidelines of one organization to another. In many cases, it also remains unclear what the considerations were how guideline developers arrived from a given level of evidence to a certain recommendation. In other cases, no specific recommendation is made. A structured format for guideline development and presentation which rates the quality of available evidence and defines the factors involved when grading the strength of the recommendations would be a powerful tool to synthesize information and could provide better guidance to practitioners.

What is GRADE and how was it developed?

The GRADE system was created in response to the need for a more unified and transparent approach to guidelines creation and reporting [7]. Individuals from all over the world, many from leading organizations involved in defining levels of evidence, including NICE, ARHQ, and the National Health and Medical Research Council (NHMRC) formed the GRADE working group in 2000 and have since been working in the development of the GRADE system [8]. This framework has now been adopted as the standard for guideline development by over 50 international organizations, including the World Health Organization (WHO), the Cochrane Collaboration, SIGN, AHRQ, and the Centers for Disease Control and Prevention (CDC). Current resources on the GRADE methodology include the original series for guideline developers published in the British Medical Journal [7–11] and the GRADE working group website [12].

How does GRADE actually work?

Just as in any well-conducted research study, a guideline should employ well-designed clinical questions which contain the four components known as "PICO": patient, intervention, comparison, and outcome of interest [13]. For example, consider a guideline which is being developed for ureteral calculi. One issue to be addressed within this guideline is the role of medical expulsive therapy. A well-designed question might ask "In adult patients with ureteral calculi, does medical expulsive therapy compared to standard care improve outcomes?" But what is the "outcome?" The GRADE system suggests attempting to identify all potentially relevant outcomes for each specific question and rate their relative importance a priori. In this example, beneficial outcomes could include stone passage rates, reduced pain, fewer complications, and fewer related surgeries. Societal outcomes could include reduced resource utilization. Physiologic outcomes could include relaxation of the ureter. Negative outcomes of the added medication could include hypotension, edema, erectile dysfunction, and increased cost. To add clarity to such a long list, GRADE ranks the relative importance of outcomes to clinical decision making on a scale of 1 (not important) to 9 (critical) [9] (Fig. 1). For all critical and most important outcomes, guideline panels using GRADE then review the available quality of evidence for each

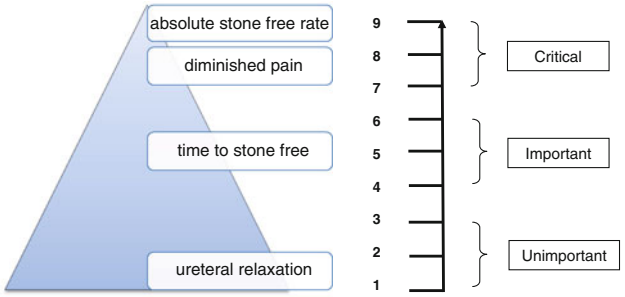

Fig. 1 Example of potential hierarchy of outcomes for medical expulsive therapy

specific outcome and list this information in an easy to interpret table [9]. Synthesis of this information into recommendations must take into consideration which outcomes are to be included in the recommendation. Should only critical (ranked 1–3) or critical plus "important" (ranked 4–6) be included? If an outcome for which evidence is of lower quality is critical for decision making, then the overall rating of quality across outcomes must reflect this lower quality evidence. If the outcome (for which evidence is of lower quality) is important but not critical, the GRADE approach suggests a rating across outcomes that reflects the higher quality evidence from the critical outcomes.

Determining the evidence quality

An important innovation of GRADE is that the quality of evidence rating is outcome specific. The background to this is that the quality of the evidence may vary across outcomes. For example, certain study limitations, such as lack of blinding of outcome assessors, may be more relevant to some outcome than others. If the outcome is an objective event such as death of any cause, then blinding may not be as important. Meanwhile, lack of blinding may bias assessments of disease-specific death. As a result, the quality of evidence for disease-specific survival may be rated lower than the quality of evidence for overall survival. An example of outcome-specific quality ratings is presented in Table 2.

Grade classifies quality of evidence into four categories: high, moderate, low, and very low (Table 3). Depending on the question being asked and the available studies, evidence may start high or low and move up or down based on predefined characteristics. Evidence from RCTs starts out as high-quality evidence but must meet strict criteria to stay at that level, while evidence from observational studies starts as low quality but may move up in certain circumstances (Table 4).

Factors which can lower quality

Limitations that may increase bias within study results include lack of allocation concealment, lack of blinding, especially for subjective outcomes, failure to adhere to intention to treat principles, large loss to follow-up, failure to report on all patient important outcomes, especially ones which would logically have been measured, and stopping early for apparent benefit. Inconsistent results (termed heterogeneity in systematic reviews), which are widely different across different studies, raise concerns that there may be true treatment differences, which cannot be explained. This becomes problematic when attempting to generalize a treatment recommendation. Indirect evidence occurs when different studies assess unique components of a question, such as results with open or laparoscopic

Table 2 Example of a theoretical evidence profile for medical expulsive therapy with alpha-blockers [25]

	Quality assessment					Summary of findings	
No. studies	Study limitation	Consistency	Directness	Precision	Publication bias	Estimate of effect	Quality
Stone free rate							
29	Serious (−1)	Moderate heterogeneity detected*	Direct	No important imprecision	Suggestion of mild publication bias	RR = 1.45; 95% CI: 1.34–1.57	++ , low
Pain reduction							
21	Serious (−1)	No important inconsistency	Indirect ≠	No important imprecision	Suggestion of mild publication bias	Analgesic requirement were lower for all studies compared to placebo	++ , low
Adverse events	Very serious (−2)	No important inconsistency	Direct	Imprecision present	Publication bias likely	Sparse reporting, range: 0–12%	+, very low

* $P = 0.01$; $I^2 = 40\%$. ≠ Many different pain assessments utilized

The I^2 statistic quantifies the relative consistency between studies on a scale of 0–100%, with higher numbers indicating more inconsistency

Table 3 Definitions for GRADE quality ratings and strength of recommendations

Quality of evidence		Strength of recommendation
High quality	++++	Strong recommendation for using an intervention
Moderate quality	+++	Weak recommendation for using an intervention
Low quality	++	Weak recommendation against using an intervention
Very low quality	+	Strong recommendation against using an intervention

Table 4 Factors that influence quality of evidence

Downgrading the evidence Quality is lowered by:	Upgrading the evidence Quality is raised by:
Limitations of study design	Large magnitude of effect
Inconsistency of results	Confounding which would reduce the effect
Indirectness of evidence	Dose–response gradient
Imprecision	
Reporting or publication bias	

surgical procedures, but not via a direct comparison. The results can then be compared indirectly, but this may lower the quality of that evidence. Commonly, this applies when different medication classes have been studied for the same outcome, but independent of one another, such as calcium channel blockers and alpha-blockers as medical expulsive therapy (MET) for ureteral stone passage [14]. Indirectness can also apply to any of the components that inform a clinical question—different patient populations, different interventions, different comparators, and different outcomes. Imprecision occurs in situations with few events that present wide confidence intervals. Even for a well-designed trial, imprecision will lower the quality rating because we are not as confident in the results. Finally, the suggestion of reporting or publication bias clouds the existing evidence with a level of uncertainty about unpublished results or studies, which also lowers the rating.

Factors which can raise quality

Observational studies cannot overcome certain bias inherent in non-randomized trials and therefore default to low-quality evidence. However, certain features can increase the quality of such studies. When methodologically strong observational studies yield large or very large and consistent estimates of the magnitude of a treatment effect, we may be confident about the results. In those situations, although the observational studies are likely to have provided an overestimate of the true effect, the weak study design is unlikely to explain all of the apparent benefit. Dose–response gradients, where increased medication doses correspond directly to increased effects, also provide stronger evidence. Finally, when potential bias would tend to oppose the effect seen, rather than enhance it, we can surmise the true effect may be even greater than what is reported. This can raise the quality rating. For example, investigators assessed treatment intensity for early-stage bladder cancer, out of concern that such intensity was not evidence based nor cost-effective [15]. Logically, more intense treatments might be expected to correlate with improved outcomes. The bias should be toward improved survival in this group, due to confounders such as more motivated and healthier patients who were willing and able to undergo more intense treatment regiments. Yet the study showed no benefit to more intense treatment. Because of the direction the bias would naturally take these results, if it were removed it would be even less likely that a benefit to intense treatment would be found. Of note, all these determinations of study quality are typically made about an entire body of evidence as represented in a systematic review and meta-analysis, rather than an individual study.

Determining the strength of a recommendation

Recommendations must always balance the desirable effects of an intervention with the undesirable effects. It is often unclear how individual patients may feel about this balance. This complexity is where GRADE attempts to provide consistency for developers and transparency for consumers. The strength of a recommendation will reflect the level of confidence we have that implementing a recommendation will do more good than harm. Many guideline systems grade their recommendations, but these may be cumbersome to interpret. The GRADE system strives for simplicity. To that end, it allows developers to ponder the quality of the evidence with three other equally important factors to derive a yes or no recommendation, with the ability to explain why the recommendation is either "strong" or "weak" [11] (Table 3). GRADE also discourages guideline developer from judgments such as "no recommendations can be made" since clinicians and patient have to make a decision. Rating the strength of the recommendation allows developers to use judgment within the rules of the GRADE framework, and the transparent presentation allows users to decide if they agree with those judgments.

Grading of Recommendations Assessment, Development, and Evaluation provides specific definitions of what

their recommendations should signal to different individuals: A strong recommendation implies that most patients would want the intervention and that physicians should routinely offer the intervention, and policy makers may adopt the practice in most situations. A weak recommendation implies that although the majority of fully informed patients would still want the intervention, a substantial proportion of patients would not and that physicians should therefore offer a discussion of alternatives, and policies may reflect the uncertainty. Best practice policy based on weak recommendations may, for example, be to ensure a discussion of options occurs.

Grading of Recommendations Assessment, Development, and Evaluation proposes four factors to determine the strength of a recommendation [11]. The first, quality of the evidence, is a key component. This rating refers to the overall quality of evidence across outcomes. The lower the quality of the evidence, the more likely a weak recommendation may be warranted. The second factor in determining recommendation strength is the level of certainty regarding the balance of advantages and disadvantages of an intervention. For an intervention with clear benefit and few side effects, such as a single dose of antibiotic to prevent urinary infections associated with certain urologic procedures [16], the decision for a strong recommendation may be clear. Even when disadvantages are significant, such as with high-dose chemotherapy for testicular cancer, a strong recommendation may reflect the fact that despite the toxicity of chemotherapy, most young men would choose to undergo this therapy for the survival advantage it affords [17]. There is little uncertainty about this balance. When uncertainty exists for benefits and harms, a weak recommendation may be appropriate. For example, older generation anticholinergics to treat overactive bladder often cause significant dry mouth and constipation [18]. The relative balance of these effects would likely differ greatly among patients, leading to large uncertainty for a recommendation.

A critical component to evidence-based medicine is incorporating patient values and preferences into clinical care [19]. This step informs the third factor in determining the strength of a recommendation in the GRADE system. When there are large variations in how different patients may value aspects of the treatment or outcomes, it is more likely that a weak recommendation may be appropriate. Such may be the case with treatments for localized prostate cancer which, for example, may have different outcomes for erectile function. It may be inappropriate to make a strong recommendation for one specific treatment over another because each treatment has a complex set of associated quality of life outcomes, which will be weighted differently by different patients [20]. A weak recommendation in this setting will prompt a discussion of options. Alternatively, patients who wish to improve the likelihood for return of continence after radical prostatectomy may be encouraged to perform pelvic floor exercises [21]. This intervention may merit a strong recommendation despite lower quality evidence because of the strong preference men in this treatment group are likely to have in favor of it and its low likelihood of harm. Any strong recommendation would also hinge on a determination that the benefits outweigh the potential harms (as outlined above).

The fourth component the GRADE system attempts to incorporate in determining the strength of a recommendation is cost. Resource allocation is becoming a vital component in healthcare policies around the world, and guideline developers may wish to take this component into careful consideration [10]. Guidelines will be used to guide policy as well as individual decision making. Cost may be more difficult to assess than other factors, especially when trying to determine broader costs like those incurred by society as a whole. Therefore, cost is best viewed with perspective and context. It is not as simple as stating that higher cost requires a weak recommendation. For example, new vaccine therapy for metastatic castrate-resistant prostate cancer is extremely expensive [22]. Policy makers would likely wish guideline developers to assume a societal perspective and factor cost into their recommendations, while individual patients with appropriate healthcare insurance may not care to be limited by cost. To what extent guideline developers consider costs varies greatly between countries; for example, the United States and the United Kingdom are at opposite extremes. While guidelines in the United States typically do not formally consider costs in the guideline development process, NICE guidelines are always accompanied by a formal cost-effectiveness analysis.

How can GRADE help urologists?

Having clear, concise, and transparent guidance on urological issues may be of benefit to any busy clinician. A physician treating a patient with a small ureteral calculus who is uncertain about the use of MET might consult the EAU/AUA Nephrolithiasis Panel's Clinical Guideline on Management of Ureteral Calculi [23]. There the physician will find this "Option": "A patient who has a newly diagnosed ureteral stone <10 mm and whose symptoms are controlled may be offered an appropriate medical therapy to facilitate stone passage during the observation period" and this "Standard": "Patients should be counseled on the attendant risks of MET including associated drug side effects and should be informed that it is administered for an "off label" use." One interpretation of these recommendations might be that the disadvantages of these medications seem to outweigh the benefits, as risk discussion is given a standard rank while usage is merely an option. Yet there is no statement that this therapy is not recommended.

What is missing is a clear way to assess the potential effect of the therapy, the quality of evidence this is based on, and what decisional factors are present which result in the recommendation being an option only. A separate assessment of the quality of evidence and an explanation of the strength of recommendation might guide the clinician in a more practical manner. For example, such an assessment performed utilizing the GRADE methodology would reference an evidence profile (Table 2) and might provide the following statement: In patients being observed for spontaneous passage of a ureteral calculus, a weak recommendation can be made for medical expulsive therapy in facilitating stone passage and reducing analgesic needs while limiting exposure to adverse events, based on low-quality evidence for passage rates and analgesic needs, and very low-quality evidence for adverse events.

How can GRADE shape our future?

The GRADE system is well designed to meet the challenges that are facing the urological community in the current era of healthcare. Comparative effectiveness research (CER) has become a distinct entity which will be relied on to inform the spectrum from clinical decision making to national health policy. A well-defined, reproducible, and consistent system that focuses on patient important outcomes but also factors in patient values and system and societal costs is ideally suited to address these challenges and provides real guidance at all levels. Quality healthcare is becoming an integral component in many systems and should be defined as a measurable outcome. The GRADE system can help identify how to measure quality for policy makers, even within the most complex decisions. For example, treatments for localized prostate cancer cannot be boiled down to a single recommendation for every patient. How can quality be measured when many complex options exist? The GRADE system can help make explicit how patient values and preferences influence the decision-making process in this disease. A recommendation for a balanced discussion of options and outcomes in this setting also identifies that process as a quality indicator. The GRADE system also provides the option to incorporate cost into the recommendations. In the United States, end-of-life care is becoming a large part of healthcare spending [24]. With many new end-of-life extending medications with large price tags on the horizon for prostate cancer, for example, Urologists will critically need a system with the ability to assess resource allocation. A structured and explicit guideline system that can be used in our globalized world would lend clarity and provide guidance. The adoption of GRADE by so many current organizations suggests it may be such a common framework.

Acknowledgments This article relies heavily on the landmark series published in the British Medical Journal by the GRADE working group.

Conflict of interest Dr. Dahm is a member of the GRADE working group.

References

1. Scottish Intercollegiate Guidelines Network (2008) SIGN 50: a guideline developer's handbook. Scottish Intercollegiate Guidelines Network, Guideline no. 50
2. Scottish Intercollegiate Guidelines Network (2004) Management of urinary incontinence in primary care. Scottish Intercollegiate Guidelines Network, Guideline no. 79
3. Winn RJ, McClure JS (2003) The NCCN clinical practice guidelines in oncology (NCCN Guidelines™) NCCN Senior Vice President, Clinical Information & Publications
4. NCCN Guidelines Prostate Cancer (2010) National comprehensive cancer network, Version 1.2011
5. Heidenreich A, Aus G, Bolla M et al (2008) EAU guidelines on prostate cancer. Eur Urol 53:68
6. Thuroff JW, Abrams P, Andersson KE et al (2011) EAU guidelines on urinary incontinence. European urol 59(3):387–400
7. Guyatt GH, Oxman AD, Vist G et al (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj 336:924
8. Schunemann HJ, Oxman AD, Brozek J et al (2008) Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. Bmj 336:1106
9. Guyatt GH, Oxman AD, Kunz R et al (2008) What is "quality of evidence" and why is it important to clinicians? Bmj 336:995
10. Guyatt GH, Oxman AD, Kunz R et al (2008) Incorporating considerations of resources use into grading recommendations. Bmj 336:1170
11. Guyatt GH, Oxman AD, Kunz R et al (2008) Going from evidence to recommendations. Bmj 336:1049
12. http://www.gradeworkinggroup.org
13. Oxman AD, Guyatt GH (1988) Guidelines for reading literature reviews. CMAJ 138:697
14. Seitz C, Liatsikos E, Porpiglia F et al (2009) Medical therapy to facilitate the passage of stones: what is the evidence? Eur Urol 56:455
15. Hollingsworth JM, Zhang Y, Krein SL et al (2010) Understanding the variation in treatment intensity among patients with early stage bladder cancer. Cancer 116:3587
16. Wolf JS Jr, Bennett CJ, Dmochowski RR et al (2008) Best practice policy statement on urologic surgery antimicrobial prophylaxis. J Urol 179:1379
17. Feldman DR, Bosl GJ, Sheinfeld J et al (2008) Medical treatment of advanced testicular cancer. JAMA 299:672
18. Meek PD, Evang SD, Tadrous M et al (2011) Overactive bladder drugs and constipation: a meta-analysis of randomized, placebo-controlled trials. Dig Dis Sci 56(1):7–18
19. Canfield SE, Dahm P (2010) Evidence-based urology in practice: incorporating patient values in evidence-based clinical decision making. BJU Int 105:4
20. Herrmann TR, Merseburger AS, Burchardt M (2009) Considerations on prostate cancer: diagnosis and treatment decisions. World J Urol 27:579
21. Hunter KF, Glazener CM, Moore KN (2007) Conservative management for postprostatectomy urinary incontinence. Cochrane Database Syst Rev (2):CD001843

22. Kantoff PW, Higano CS, Shore ND et al (2010) Sipuleucel-T immunotherapy for castration-resistant prostate cancer. N Engl J Med 363:411
23. Preminger GM, Tiselius HG, Assimos DG et al (2007) Guideline for the management of ureteral calculi. J Urol 178:2418
24. Emanuel EJ (1996) Cost savings at the end of life. What do the data show? JAMA 275:1907
25. Seitz C, Liatsikos E, Porpiglia F et al (2009) Medical therapy to facilitate the passage of stones: what is the evidence? Eur Urol 56:455