Grading Quality of Evidence and Strength of Recommendations for Diagnostic Tests and Strategies

Andrea Rita Horvath1*

Faced with the plethora of new diagnostic and therapeutic interventions, busy physicians need clear guidance on the best approaches to follow for their patients. This need has led to such a proliferation of practice guidelines (PGs) that for diabetes mellitus alone, for example, more than 150 guidelines are available worldwide. In the “jungle” of PGs, many provide conflicting guidance, and the literature displays extensive variation in the approaches to formulating recommendations. Therefore, there is an international move toward standardizing guideline methodology so that recommendations are conceived in a systematic and transparent process and that the link between the evidence and the strength of recommendations is explicitly documented. This commentary provides a brief overview and the strength of recommendations is explicitly documented. This commentary provides a brief overview of the principles for assessing the strength of evidence and the challenges guideline developers face in formulating graded recommendations related to the use of laboratory tests.

Guidelines aim to close the gap between research and practice and to provide rigorously developed, valid, and applicable recommendations for achieving the best possible outcomes. The formulation of evidence-based guidelines implies a process in which the body of evidence has been systematically explored, its quality critically evaluated, and the research findings synthesized and translated into recommendations for best practice. In PGs, quality of evidence indicates the degree of confidence that the evidence is adequate to support recommendations. Quality of evidence can be judged by considering the following aspects (1):

1. Study design usually defines the level of evidence. For example, questions on the efficacy of treatment are best answered by randomized controlled trials (RCTs), and questions about diagnostic accuracy are best addressed by properly designed prospective cohort studies.
2. Internal validity refers to a lack of design-related biases that could threaten the soundness of the study. In diagnostic accuracy studies, various forms of verification biases, spectrum bias, or review bias can lead to overestimates of diagnostic performance.
3. Consistency refers to the similarity of estimates of effect across studies.
4. Directness is the extent to which the study’s patients, interventions, and outcomes are similar to those in practice. Diagnostic accuracy is a surrogate for important outcomes for patients and thus is considered to provide indirect evidence (2).
5. Precision refers to the reliability of an estimate of effect and is best described by the width of the 95% CI. Precision is influenced by the sample size of the study, the techniques for measuring the analyte, and the variation in analyte values in the population.
6. Other factors, such as reporting bias, can lower the quality of the evidence, whereas a strong association or the presence of a dose–response gradient can increase the quality of the evidence (1).

Beyond scientific judgment of the quality and strength of evidence, guideline developers need to make value judgments before formulating final recommendations. In this phase, most PGs assess the strength of recommendations. The latter indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm (1). Value judgments about the strength of a recommendation imply that, in addition to evidence, guideline developers have given due consideration to the following practical aspects: balance between benefits and detriments; transferability of the evidence to the given population, condition, or outcomes; preferences of the patient; impact on healthcare organization; and costs (1).

The process of considered judgment introduces subjectivity into the interpretation and translation of the evidence. To avoid the dominance of opinion of individual experts or professions, it is advised that this phase be carried out by a multidisciplinary team and that the process be well documented, including the reporting of potential conflicts of interests. To this end, most grading systems rate separately the quality of a body of evidence and the strength of recommenda-
tions. It is assumed that graded recommendations that are developed in a more rigorous and transparent fashion will raise public confidence in the process and lead to better medical decisions and improved patient outcomes.

On the basis of these assumptions, the US Congress mandated the Agency for Healthcare Research and Quality (AHRQ) to evaluate schemes for grading the strength of evidence underlying recommendations. Of the 121 critically evaluated schemes, only 19 for assessing study quality and 7 for rating strength of evidence met the AHRQ quality criteria (3). The Canadian Optimal Medication Prescribing and Utilization Service updated the AHRQ review and recommended only 2 grading systems for use (4). One of these is the GRADE approach, which offers a uniform system for grading all types of recommendations, including diagnostic ones (1, 2). GRADE is rapidly becoming a standard and already has been adopted by many journals and guideline agencies.

For PGs in which scope was targeted mostly at the laboratory management of certain conditions, we found a great variety of approaches. For example, an examination of the grading schemes used in the 10 most recently published National Academy of Clinical Biochemistry (NACB) guidelines (http://www.nacb.org) revealed 4 main categories of approaches:

1. PGs that adopted the same grading system as used by the relevant clinical society with and for whom the guideline was developed (e.g., the guidelines on hepatic injury, on acute coronary syndrome and cardiovascular risk factors, and on diabetes mellitus);
2. PGs that adopted the grading system specifically developed for NACB (e.g., the point-of-care testing, pharmacogenetics, and newborn-screening guidelines);
3. PGs that developed their own grading system (e.g., the tumor marker guideline); and
4. PGs that were consensus based (e.g., on toxicology) or used no grading whatsoever (e.g., the thyroid guideline).

The question of whether diagnostic recommendations need a different grading system than therapeutic ones is controversial. The reasoning of guideline groups that favor the approach in the first category in the above list is that if a PG is developed in a multidisciplinary fashion with and for clinicians, then it is advisable to use a common system that is familiar to those who will use the recommendations. Although such reasoning is well justified, the problem with this approach is that most of the grading systems adopted by clinical societies were developed primarily for therapeutic recommendations and thus use a scale that views multiple RCTs as the highest category of evidence for demonstrating improved patient outcomes. Application of such systems to grading recommendations related to diagnostic or prognostic questions presents a great challenge. Most guideline panels are encouraged to formulate recommendations that will directly affect outcomes important to the patient; however, most recommendations in diagnostic guidelines address issues of “when and how a test should be used to rule in or out a diagnosis or to monitor a disease,” or the prognostic value of the tests. In these types of questions, the surrogate outcome is diagnostic or prognostic accuracy, the evidence for which commonly derives from well-designed cohort studies rather than RCTs. Consequently, most grading systems currently in use downgrade the quality of evidence right at the start, because the evidence is not from RCTs.

Furthermore, clinicians diagnose conditions by means of a battery of clinical, laboratory, and imaging tests or testing strategies to decide on the best options for managing their patients. The patient’s final outcome will be influenced by whether the clinician, with the help of tests, correctly identifies the ailment and whether the treatments that he/she chooses to administer throughout the course of the disease are the most effective for that particular patient and do not cause more harm than good. In other words, improved outcomes for a patient will be reflected in the net benefit of all interventions given to the patient. Judging the contribution of laboratory testing to this net benefit is not always straightforward, because test results or test accuracy are often surrogates for hard measures of patient-centered outcomes, such as morbidity or mortality. The GRADE group acknowledges these difficulties and limits the use of its system to scenarios “when diagnostic intervention studies—ideally randomised controlled trials but also observational studies—comparing alternative diagnostic strategies with assessment of direct patient-important outcomes are available.” When such studies are not available, as is frequently the case, the GRADE group advises that guideline panels focus on studies of test accuracy and make inferences about the likely impact on patient-important outcomes (2).

The aim of grading recommendations is to promote the use of evidence-based approaches in healthcare and to provide guideline groups with a tool for making more informative recommendations; however, the wealth of varying grading schemes currently in use for PGs threatens to produce not only widely divergent approaches but also confusion, and even resistance to evidence-based medicine. To avoid the confusion and unnecessary variations in practice caused by differing interpretations and judgments of the quality and value of the evidence base informing PGs and to facilitate international accord about what constitutes an evidence-based recommendation, guideline agen-
cies and editors of medical journals need to arrive at collective agreement on reporting standards for guidelines and a standardized nomenclature of evidence hierarchies and grades of recommendations (5).

Until such standards of guideline reporting and grading become universally accepted for diagnostic recommendations as well and until teams gain the necessary skills to apply these methods, we advise PG panels to formulate their recommendations in a transparent evidence- and consensus-based process of systematically reviewing, critically appraising, and discussing the evidence, including highlighting the pros and cons of the evidence and specifically addressing the gaps in the knowledge base behind the recommendations.

**Author Contributions:** All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

**Authors’ Disclosures of Potential Conflicts of Interest:** Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:

**Employment or Leadership:** None declared.

**Consultant or Advisory Role:** A.R. Horvath, member, international advisory board, Becton Dickinson.

**Stock Ownership:** None declared.

**Honoraria:** A.R. Horvath, Becton Dickinson.

**Research Funding:** None declared.

**Expert Testimony:** None declared.

**Role of Sponsor:** The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

**Acknowledgments:** The author thanks Jako Burgers (Dutch Institute for Healthcare Improvement CBO, the Netherlands), and Jan Brozek, Paul Glasziou, and Holger Schünemann (for the GRADE group) for scientific discussions and valuable suggestions for the manuscript.

**References**