Introducing GRADE: a systematic approach to rating evidence in systematic reviews and to guideline development

Marcel Dijkers, PhD, FACRM
Icahn School of Medicine at Mount Sinai, Dept. of Rehabilitation Medicine

Dr. Marcel Dijkers, rehabilitation researcher at the Icahn School of Medicine at Mount Sinai, presents another in a series of brief articles around evidence-based research and knowledge translation topics. This article explains the GRADE process (Grading of Recommendations Assessment, Development and Evaluation) and explores its usefulness for rehabilitation and disability research.

GRADE (Grading of Recommendations Assessment, Development and Evaluation) is a well-developed formal process to rate the quality of scientific evidence in systematic reviews and to develop recommendations in guidelines that are as evidence-based as possible. GRADE was developed by an international panel, including members of some of the premier evidence-based practice centers (McMaster, Harvard, the Norwegian and German Cochrane Centres, etc.). While there were some earlier publications,1-7 a series of papers published in the Journal of Clinical Epidemiology from 2011 to 2013 constitute the most complete and systematic expose.8-22 More information can be found on the GRADE Working Group’s website (http://www.gradeworkinggroup.org).

A number of panels and agencies have adopted GRADE, among others the Cochrane Collaboration (the Effective Practice and Organisation of Care group, the Public Health and other groups), World Health Organization (various guideline development groups), England’s National Institute for Health and Clinical Excellence (NICE); the Canadian Task Force on Preventive Health Care, the Norwegian Knowledge Centre for the Health Services, the CDC’s Advisory Committee on Immunization Practices, the Kaiser Permanente National Guideline Program, and some groups in the Campbell Collaboration. Some use it with modifications (not recommended by GRADE), and some report minor or major challenges in using the GRADE process unmodified.23

GRADE was designed for reviews and guidelines that examine alternative clinical management strategies or interventions, which may include no intervention or current best management. In developing GRADE, the authors considered a wide range of clinical questions, including diagnosis, screening, prevention, and therapy. For that reason, the system can also be applied to rehabilitation, public health, and health systems questions.
GRADE is much more than a rating system, such as those published by various Evidence-Based Practice (EBP) organizations. It offers a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines and for carrying out the steps involved in developing recommendations. GRADE specifies an approach to framing questions, choosing outcomes of interest and rating their importance, evaluating the evidence, including making explicit the risk of various biases, and taking into account issues of imprecision (i.e. broad confidence intervals), inconsistency of results between studies, and indirectness (i.e. using evidence from a similar population, e.g. stroke instead of traumatic brain injury).

GRADE incorporates evidence with explicit consideration of the values and preferences of patients and society at large to arrive at recommendations. Furthermore, it provides clinicians and patients/clients with a guide to using those recommendations in clinical practice, and policy makers with a guide to their use in health policy.

Based on the recommendations by Johnston and Dijkers for improved evidence standards, a review of a number of existing approaches to systematic reviewing of evidence and developing recommendations (Cochrane Collaboration; Campbell Collaboration, American Academy of Neurology, Centre for Evidence-Based Medicine, among others) likely comes to the conclusion that GRADE is the most flexible methodology with respect to evaluating the evidence (downgrading, upgrading, handling indirect evidence, etc.). It also goes beyond the other systems where it concerns the translation of evidence into recommendations. The special emphasis in GRADE on the values and preferences of consumers (which now is being adopted by others) fits eminently with the traditional emphasis in rehabilitation and disability studies.

The GRADE methodology is applicable whether the quality of the relevant evidence is high or low. The GRADE system was among the first to lay out a systematic way of evaluating whether evidence should be downgraded—for instance, a randomized controlled trial (RCT) executed with poor allocation concealment and high attrition should not be considered to be equivalent to a well-done RCT (see Table 1).

In addition, GRADE was the first to specify under what circumstances evidence from a study may be upgraded—for instance, when an effect size is very large, a dose-response gradient is shown, or other circumstances would suggest that what traditionally has been considered a “rather weak design” (e.g. a case-control study) may produce evidence that is of a level produced by an average RCT.
Table 1. Factors that may lead to downgrading or upgrading of evidence in the GRADE approach

<table>
<thead>
<tr>
<th>Downgrading</th>
<th>Upgrading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Serious risk of bias</td>
<td>1. Large effect size</td>
</tr>
<tr>
<td>2. Serious inconsistency</td>
<td>2. Dose-response gradient</td>
</tr>
<tr>
<td>3. Serious indirectness</td>
<td>3. All plausible confounding would reduce a demonstrated effect</td>
</tr>
<tr>
<td>4. Serious imprecision</td>
<td>4. All possible confounding would suggest a spurious effect when the actual results show no effect</td>
</tr>
<tr>
<td>5. Likely publication bias</td>
<td></td>
</tr>
</tbody>
</table>

Another advantage of GRADE is that it requires the systematic reviewer to make explicit his or her judgment of each factor that determines the quality of evidence for each outcome. Because alternative diagnostic or therapeutic approaches may all have a balance of positive and negative outcomes (costs, side effects, positive effects in various domains), a guideline developer needs to find a way to systematically identify these, and weigh evidence for all of them simultaneously in making recommendations; GRADE offers a systematic approach to resource use\(^\text{17}\) and to handling multiple outcomes.\(^\text{18}\) Finally, a computer program (GRADEpro) with its associated help file facilitates the development of evidence tables (in GRADE called evidence profiles, or EPs\(^\text{19,20}\)) and summary of findings (SoFs) tables that are based on the EPs.

Figure 1 presents a schematic view of GRADE’s process for developing recommendations; the top half describes steps in the process common to systematic reviews and to guideline development, and the lower half describe steps that are specific to guideline creation. One begins by defining the question in terms of the populations, alternative management strategies (an intervention, sometimes experimental and a comparator, sometimes standard care), and all patient-important outcomes (in this case three). The authors have provided guidance as to which clinical and other questions are suitable for answering with GRADE (or with any systematic review approach, for that matter) and for collecting evidence.
Figure 1. Overview of the GRADE process (modified from Guyatt et al.\textsuperscript{5})

- Question (clinical, public health, health system, etc.)
  - Study 1
  - Study 2
  - Study 3
  - Study 4
  - Study 5
    - Outcome 1
    - Outcome 2
    - Outcome 3
      - Important outcomes
      - Critical outcome

**Generate an estimate of effect for each outcome.**

**Rate the quality of the evidence for each outcome, across studies.**
- Reduce the rating as needed (study limitations, imprecision, inconsistency of results, indirectness of evidence, likelihood of publication bias).
- Increase the rating as warranted (large effect size, dose-response curve, confounders are likely to minimize the effect).

**Final rating of evidence quality for each outcome:** high, moderate, low or very low.

**Rate the overall quality of the evidence.**
(lowest quality among critical outcomes)

**Decide on the direction (for/against) and grade strength (strong/weak) of the recommendation, considering:**
- Quality of the evidence
- Balance of desirable/undesirable outcomes
- Values and preferences of patients/clients/others

**Decide if any revision of direction or strength is necessary considering:** resource use.
For guidelines, one classifies the outcomes as either critical (one outcome in Figure 1) or important but not critical (two outcomes). A systematic search leads to inclusion of relevant studies (in this scheme, five such studies). Systematic reviewers or guideline authors then use the data from the individual eligible studies to generate a best estimate of the effect on each patient-important outcome and an index, typically a confidence interval (CI), of the uncertainty associated with that estimate.

The Figure illustrates that evidence must be summarized for each patient-important outcome—the summaries ideally coming from optimally conducted systematic reviews. For each comparison of alternative management strategies, all outcomes should be presented together in one EP or SoFs table. It is likely that all studies relevant to a rehabilitation or disability question will not provide evidence regarding every outcome. For example, Figure 1 shows the first study providing evidence for the first and second outcome, the third study for outcomes two and three, and so on. Indeed, there may be no overlap between studies providing evidence for one outcome and those providing evidence for another. For instance, RCTs may provide the relevant evidence for benefits, and observational studies provide the evidence for rare but serious adverse effects.

In the GRADE approach, RCTs start as high-quality evidence and observational studies as low-quality evidence to support estimates of intervention effects. As described above, five factors may lead to rating down the quality of evidence and three factors may lead to rating up (see Table 1). Ultimately, the quality of evidence for each outcome falls into one of four categories from high to very low. Systematic review and guideline authors use this approach to rate the quality of evidence for each outcome across studies (i.e., for a body of evidence). This does not mean rating each study as a single unit. Rather, GRADE is “outcome centric” in that a rating is made for each outcome, and quality may differ—indeed, is likely to differ—from one outcome to another within a single study and across a body of evidence.

Guideline developers (but not systematic reviewers) then review all the information to make a decision about which outcomes are critical and which are important, and come to a final decision regarding the rating of the overall quality of evidence. They next need to consider the direction and strength of recommendation. The balance between desirable and undesirable outcomes and the application of patients’ values and preferences determine the direction of the recommendation; these same factors, along with the quality of the evidence, determine the strength of the recommendation. Both direction and strength may be modified after taking into account the resource use implications of the alternative management strategies.22
Because most existing systematic reviews do not adequately address all relevant outcomes in a single document, the GRADE process may require relying on more than one systematic review. Systematic reviews often address more than one comparison. They may evaluate an intervention in two disparate populations or examine the effects of a number of interventions. Such reviews are likely to require more than one SoFs table. For example, a review of cognitive remediation may evaluate the effectiveness of training of executive function for different populations, such as those with mild versus moderate traumatic brain injury.

GRADE has been used for a number of systematic reviews and guidelines, including quite a few that would be of interest to professionals with an interest in rehabilitation and disability. Examples: nonpharmacologic interventions for osteoarthritis\(^2\) and for spasticity in Multiple Sclerosis (MS),\(^2\) rehabilitation interventions for nonambulatory MS patients;\(^2\) music interventions for psychological and physical outcomes in cancer;\(^2\) art therapy for psychosomatic disorders;\(^2\) exercises\(^3\) and manipulation\(^3\) for chronic neck pain; interventions for depression in the workplace;\(^3\) parent interventions for children with intellectual disabilities;\(^3\) childhood adversity as a cause of schizophrenia;\(^3\) family-based cognitive-behavioral therapy for children and adolescents with obsessive-compulsive disorder;\(^3\) the benefits of physical activity for youth with developmental disability\(^3\) and interventions to enhance return to work by cancer patients\(^3\)—quite a variety. It may be worth your while to investigate whether GRADE is making the grade.

References


Suggested citation:

http://www.ktdrr.org/products/update/v1n5/dijkers_grade_ktupdatev1n5.html

*KT Update* was developed by the Center on Knowledge Translation for Disability and Rehabilitation Research (KTDRR) through grant #H133A120012 to SEDL from the National Institute on Disability and Rehabilitation Research (NIDRR) in the US Department of Education's Office of Special Education and Rehabilitative Services (OSERS). However, these contents do not necessarily represent the policy of the US Department of Education, and you should not assume endorsement by the federal government.

SEDL operates the Center on KTDRR. SEDL is an Equal Employment Opportunity/Affirmative Action Employer and is committed to affording equal access to education and employment opportunities for all individuals. SEDL does not discriminate on the basis of age, sex, race, color, creed, religion, national origin, sexual orientation, marital or veteran status, or the presence of a disability.

Copyright © 2013 by SEDL.