The medical community relies on scientific evidence to guide clinical practice. Evidence from systematic reviews, randomized controlled clinical trials (RCTs), case–control or cohort studies, observational studies, and expert opinions are used to make disease-specific practice recommendations. More than 100 grading systems are used to rate the strength of these recommendations (West et al., 2002). A centralized and transparent method for evaluating and comparing these studies with the goal of translating evidence-based medicine to clinical practice guidelines is the cornerstone of two such validation scales: the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) and Strength of Recommendation Taxonomy (SORT).

In the GRADE system, one frames a question, chooses critical and important outcomes by which to judge the existing body of evidence, rates the quality for each outcome, and finally decides on the direction (for or against) and strength (strong or weak) for the recommendation considered. The SORT method is a simpler rating scale that judges the study quality and strength of recommendation based on patient-oriented evidence (Table 1). Whereas GRADE and SORT evaluate the body of evidence to establish sound guidelines, the AGREE (Appraisal of Guidelines Research and Evaluation) instrument provides a framework for assessing the quality of development of clinical practice guidelines. AGREE is a generic instrument that provides an assessment of the validity and likelihood that the stated guideline will achieve its outcome.

**ADVANTAGES AND LIMITATIONS**

**GRADE**
- Adopted by more than 65 organizations worldwide as an international standard for guideline development.
- Explicitly evaluates relevant outcomes and considers risks, benefits, patient expectations, and resource utilization in reaching recommendations.
- Using GRADE is difficult and requires expertise in statistics.

**SORT**
- Adopted by the American Academy of Dermatology.
- Simple and easily applied by authors and physicians.
- Advocates the use of patient-oriented rather than disease-oriented outcomes.
- Overly simplified instrument that is not applied internationally.

**AGREE**
- Validated instrument that assesses the quality of guideline development.
- Assesses six domains in guideline development: scope and purpose, stakeholder involvement, rigor of development, clarity and presentation, applicability, and editorial independence.

The GRADE process begins with asking a clinically relevant, well-designed clinical question composed of four elements: a patient, problem, or population; an intervention; a comparison intervention; and an outcome. The second step in the GRADE system is to gather the best evidence to answer the question. The third step is assessing the quality of evidence and the confidence in the estimates of the treatment.
The fourth step evaluates the trade-off between risks and benefits, reflecting the best assessment of patients’ perspective of the evidence before making the final recommendation (Guyatt et al., 2013) (Figure 1).

The study design determines the initial quality of evidence rating. RCTs start as high-quality evidence, whereas observational studies begin as low-quality evidence. This ranking can be upgraded or downgraded based on specific factors that can affect the quality of evidence. Factors that can lower the quality of evidence include study limitations, inconsistencies in the results, indirectness of evidence, imprecision in the estimates, and publication bias. The rating can be upgraded if the study shows the presence of a dose-response effect or a large magnitude of the estimated effect.

After assessing all the domains, the body of evidence per outcome is categorized as high (+++), moderate (++), low (+), or very low (+) (Mustafa et al., 2013). The quality of evidence rating is summarized in the Evidence Profile (EP) table, which includes an explicit judgment of each factor that determines the quality of evidence. Table 2 is an example of a transparent and concise way of showing the guideline panel judgments about the domains. It also contains the Summary of Findings Table (SoF). The SoF is a quantitative assessment of the confidence in the estimates of effects (i.e., relative risk), without a qualitative judgment of the evidence rating that is provided in the EP table. The EP and the SoF tables serve different purposes and are directed toward different audiences. EP are intended for review authors and anyone who questions a quality of assessment. SoF are designated for a broader audience, such as users of systematic review and guidelines (Guyatt et al., 2011).

The fourth step of the process is assessing the values and preferences of the target population regarding their beliefs and expectations for their health and life. This step refers to the process in which individuals weigh the potential benefits, harms, costs, limitations, and inconveniences of treatment options in relation to one another. With this information, the panel is more equipped to accurately define the trade-off between the benefits (desirable outcome) and risks (undesirable consequences) for a particular intervention. Ideally, “the panel” (guideline developers) will conduct a systematic review summarizing relevant studies regarding the patient’s values and preferences. The greater the variability or uncertainty in values and preferences, the more likely a weak recommendation is warranted (Andrews et al., 2013).

The overall strength of recommendation is based on the balance of risks and benefits, the quality of evidence, the values and preferences of the patients, and costs required for the treatment. Each component is given equal weight in relation to the other components. This strength of recommendation ranges on a continuum of categories from “strongly for” to “strongly against” the intervention (Table 1). If the panel is highly confident of the balance between desirable and undesirable consequences, they make a strong recommendation for (desirable outweighs undesirable) or against (undesirable outweighs desirable) an intervention (Andrews et al., 2013).

Guideline panels may also choose to make special recommendations when there is insufficient evidence, for example, an “only-in-research” recommendation. This recommendation is used when further research may reduce uncertainty.

---

**Table 1. Comparison between GRADE and SORT with regard to the strength of recommendation and the quality of evidence**

<table>
<thead>
<tr>
<th>Strength of recommendation</th>
<th>Quality of the evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE</td>
<td></td>
</tr>
<tr>
<td>Strong for = benefits outweigh risks of the intervention</td>
<td>High quality = further research is very unlikely to change our confidence in the estimate of effect</td>
</tr>
<tr>
<td>Strong against = risks outweigh benefits of the intervention</td>
<td>Moderate quality = further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate</td>
</tr>
<tr>
<td>Weak = most informed people would choose this recommendation but a substantial number would not (risks and burdens finely balanced)</td>
<td>Low quality = further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate</td>
</tr>
<tr>
<td></td>
<td>Very low quality = any estimate is very uncertain</td>
</tr>
<tr>
<td>SORT</td>
<td></td>
</tr>
<tr>
<td>A = based on consistent and good quality patient-oriented evidence</td>
<td>Level 1 = good quality, patient-oriented</td>
</tr>
<tr>
<td>B = based on inconsistent or limited quality patient-oriented evidence</td>
<td>Level 2 = limited quality, patient-oriented</td>
</tr>
<tr>
<td>C = based on consensus, usual practice, opinion, disease-oriented evidence or case series</td>
<td>Level 3 = other evidence (usual practice, opinion, disease oriented evidence)</td>
</tr>
</tbody>
</table>
about the intervention and further research is considered of
good value for the anticipated costs. Alternatively, the panel
may decide not to make recommendations for or against a par-
ticular strategy if they find the strength in the estimate is too
low, the trade-off between risks and benefits is too close, or
values, preferences, and resource implications are not known
(Andrews et al., 2013). The main limitation for using GRADE is
that it is a complex methodology with a steep learning curve.

STRENGTH OF RECOMMENDATION TAXONOMY (SORT)
SORT was developed by the editors of U.S. Family Medicine
and Primary Care journals and the Family Practice Inquiries
Network as an initiative to construct a unified taxonomy that
allows authors to rate individual studies or bodies of evi-
dence (Ebell et al., 2004). The SORT approach is the main
methodology that the American Academy of Dermatology
utilizes in its guideline development process.

The SORT process addresses the quality, quantity, and
consistency of evidence, and it emphasizes the use of
patient-oriented outcomes that measure changes in morbid-
ity or mortality (Ebell et al., 2004). The expert panel reviews
the bodies of evidence for each of the recommendations and
assigns a strength of recommendation on a scale of A through

![SORT Diagram]

Table 2. Evidence to recommendation framework

<p>| Question/recommendation: Should pulmonary rehabilitation vs. usual community care be used for COPD with recent exacerbation? |
|---|---|---|---|
| Population: Patients with COPD and recent exacerbation of their disease |
| Intervention: Pulmonary rehabilitation vs. no rehabilitation |
| Setting (if relevant): Outpatient |</p>
<table>
<thead>
<tr>
<th>Decision domain</th>
<th>Judgment</th>
<th>Reason for judgment</th>
<th>Subdomains influencing judgment</th>
</tr>
</thead>
</table>
| Balance of desirable and undesirable outcomes | Yes | The desirable consequences are substantial (including substantial reduction in hospitalization, small but important reduction in mortality, and improvement in quality of life that exceeds the minimal important difference) and valued highly. The undesirable consequences, inconvenience, and burden are relatively minor and associated with minimal disability. | Baseline risk for desirable and undesirable outcomes: 
- Is the baseline risk similar across subgroups? 
- Should there be separate recommendations for subgroups? 
- Relative risk for benefits and harms: 
  - Are the relative benefits large? 
  - Are the relative harms large? 
- Requirement for modeling: 
  - Is there a lot of extrapolation and modeling required for these outcomes? 
  - Typical values: 
    - What are the typical values? 
    - Are there differences in the relative value of the critical outcomes? 
- Confidence in estimates of benefits and downsides, confidence in estimates of resource use. Consider all critical outcomes, including the possibility that some may not be measured. 
  - Key reasons for rating evidence down or rating up |
| Given the best estimate of typical values and preferences, are you confident that the benefits outweigh the harms and burden or vice versa? | | | |
| Confidence in estimates of effect (quality of evidence) | Yes | There is moderate-mortality, function, and quality-of-life outcome-to-high hospitalizations) quality evidence for the desirable consequences, and quality evidence for the undesirable (burden) | Source of typical values (panel or study of general population or patients) 
Source of estimates of variability and extent of variability 
Method for determining values satisfactory for this recommendation |
| Is there high or moderate quality evidence? | | | |
| Values and preferences | Yes | We can be confident that patients place a high value on avoiding hospitalizations and mortality as well as improving quality of life and a low value on avoiding the inconvenience associated with rehabilitation. We can be confident that these values vary little among patients with chronic respiratory disease. | Source of typical values (panel or study of general population or patients) 
Source of estimates of variability and extent of variability 
Method for determining values satisfactory for this recommendation |
| Are you confident about the typical values and preferences and are they similar across the target population? | | | |
| Resource implications | Yes | There are resources required to provide pulmonary rehabilitation but these are balanced by decreased resource needs as a result of decreased hospitalizations and net cost is well worth it given the desirable outcomes. | What are the costs per resource unit? 
Feasibility: 
- Is this intervention generally available? 
Opportunity cost: 
- Is this intervention and its effects worth withdrawing or not allocating resources from other interventions 
Differences across settings: 
- Is there lots of variability in resource requirements across settings? |
| Are the resources worth the expected net benefit from following the recommendation? | | | |
| Overall strength of recommendation | Strong | The guideline panel recommends that patients with recent exacerbations of their COPD undergo pulmonary rehabilitation (Note: this is a hypothetical recommendation developed for this article and not intended for clinical decision making). | |
| Evidence to recommendation synthesis | | The moderate-to-high confidence in the moderate-to-large magnitude of effects on highly valued outcomes, and the moderate-to-high confidence that undesirable outcomes are modest and their avoidance not highly valued suggest a strong recommendation. | |

COPD, chronic obstructive pulmonary disease.
Reprinted with permission from Andrews et al., 2013.
C. For example, consistent and good-quality evidence for treatment at an A-level rating would include a systematic review/meta-analysis with consistent results or a high-quality, large individual RCT.

An A-level recommendation is based on consistent and good-quality, patient-oriented evidence. A B-level recommendation is based on inconsistent or limited-quality patient-oriented evidence. A C-level recommendation is based on consensus, usual practice, opinion, disease-oriented evidence, or case series for studies of diagnosis, treatment, prevention, or screening (Table 1). The main limitation of SORT is that it is an overly simplified instrument that is not applied internationally.

APPRaisal OF GUIDELINES RESEARCH AND EVALUATION (AGREE)

Whereas GRADE and SORT evaluate the body of evidence to establish sound guidelines, the AGREE instrument assesses the quality of the development of clinical practice guidelines. The quality of guidelines is based on the confidence that potential biases have been addressed adequately, that recommendations are both internally and externally valid, and that they are feasible for practice. New or existing guidelines and updates of existing guidelines may be appraised with AGREE. It is a validated tool with a 4-point numerical scoring system, ranging from 1 (representing strongly disagree) to 4 (strongly agree). Scores reflecting inadequate quality are assigned a score ≤2. This instrument can be applied to any disease area, including those in diagnosis, health promotion, and treatment.

AGREE is composed of 23 key items encompassed within six domains. Each domain is intended to capture a different dimension of the guideline quality: scope and purpose, stakeholder involvement, rigor of development, clarity and presentation, applicability, and editorial independence. The domain score is calculated by adding all of the individual item scores in a domain and standardizing the total as a percentage of the maximum possible score for that domain. Each domain score may be useful for comparing guidelines and will aid in the decision whether to use that guideline. There is no set threshold for the domain score by which to define a “good” or “bad” guideline. Finally, an overall assessment is made as to the quality of the guideline, taking each of the appraisal criteria into account and rating it as “strongly recommend,” “recommend (with provisos or alteration),” “would not recommend,” or “unsure” (AGREE Collaboration, 2001).

Recently, AGREE was modified to AGREE II. The purpose of this updated version was to improve reliability, validity, and supporting documentation. The newer version continues to have 23 items and six domains, whereas the rating scale for each domain has become more detailed, using a 7-point rather than 4-point scale. Score 1 is assigned when there is no relevant information; scores between 2 and 6 are given when the domain does not meet the full criteria; and a maximum score of 7 is given to exceptional reports (AGREE Next Steps Consortium, 2009).

QUESTIONS

This article has been approved for 1 hour of Category 1 CME credit. To take the quiz, with or without CME credit, follow the link under the “CME ACCREDITATION” heading.

1. Factors that may decrease the strength of evidence for randomized controlled trials (RCTs) in GRADE include all of the following except
A. High likelihood of publication bias.
B. Inconsistency.
C. Large or very large treatment effect.
D. Indirectness of evidence.

2. Which of the following is false regarding AGREE?
A. It is a method of grading the strength of a recommendation.
B. It can be applied to any disease of interest.
C. It is an instrument that assesses the quality of the development of clinical practice guidelines.
D. It has no set threshold for the domain score by which to define a “good” or “bad” guideline.

3. Which of the following is false about GRADE?
A. It has been adopted by more than 65 organizations worldwide.
B. It was developed by an international group composed of guidelines developers, systematic reviewers, clinicians, public health officers, researchers, methodologists, and other health professionals.
C. It explicitly evaluates outcomes and sets forth recommendations.
D. It assigns strength of recommendations on a scale of A through C.

4. Regarding the SORT process, all of the following are true except
A. It is a method for assessing patient-oriented versus disease-oriented evidence.
B. It is a complicated methodology that may be a barrier for widespread use.
C. It is simple and easily applied to daily practice by authors and physicians.
D. It assigns a strength of recommendations on a scale of A through C and quality of the evidence on levels 1 through 3.

5. The Evidence Profile table
A. Provides a qualitative judgment of the evidence rating.
B. Is directed toward users of systematic reviews.
C. Is part of the SORT process.
D. Summarizes the values and preferences of the target population.
SUMMARY

The AGREE instrument has been applied towards the critical appraisal of clinical practice guidelines and adaptation in evidence-based guidelines for “prevention of skin cancer” by the German Guideline Program in Oncology. The rigorous inclusion criteria required by the AGREE instrument narrows the 480 citations related to the topic “prevention of skin cancer” to only 12 studies. The strict criteria needed to be fulfilled by the AGREE tool demonstrate that methodological flaws are an important obstacle in the development of practical guidelines (Petrarca et al., 2013). The AGREE instrument was also chosen as the appraisal tool for evaluation of quality of clinical practical guidelines for treatment of psoriasis vulgaris, 2006-2009 (Tan et al., 2010).

GRADE and SORT are two methods of evaluating a body of evidence and the quality of studies to create a comprehensive recommendation. The AGREE instrument is a validated quantitative scoring method created to systematically assess the quality of practice guidelines. Knowledge of these commonly applied grading systems is important for the informed dermatologist and clinician to understand for clinical practice and guideline development.

CONFLICT OF INTEREST

The authors state no conflict of interest.

CME ACCREDITATION

This CME activity has been planned and implemented in accordance with the Essential Areas and Policies of the Accreditation Council for Continuing Medical Education through the Joint Sponsorship of ScientiaCME and Educational Review Systems. ScientiaCME is accredited by the ACCME to provide continuing medical education for physicians. ScientiaCME designates this educational activity for a maximum of one (1) AMA PRA Category 1 Credit. Physicians should claim only credit commensurate with the extent of their participation in the activity.

To take the online quiz, follow the link below:
http://continuingeducation.dcri.duke.edu/research-techniques-made-simple-journal-based-cme-1

SUPPLEMENTARY MATERIAL

A PowerPoint slide presentation appropriate for teaching purposes is available at http://dx.doi.org/10.1038/jid.2014.335.

REFERENCES


