Grading Diagnostic Evidence-Based Statements and Recommendations

Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group

www.gradeworkinggroup.org

Acknowledgement

Jeff Andrews, associate professor^x
David Atkins, chief medical officer^a
Dana Best, assistant professor^b
Peter A Briss, chief^c
Martin Eccles, professor^d
Yngve Falck-Ytter, associate director^e
Signe Flottorp, researcher^f
*Gordon H Guyatt, professor^g
Robin T Harbour, quality and information director^h
Margaret C Haugh, methodologist^i
David Henry, professor^j
Suzanne Hill, senior lecturer^j
Roman Jaeschke, clinical professor^k
Gillian Leng, guidelines programme director^l
Alessandro Liberati, professor^m
Nicola Magrini, director^n
James Mason, professor^d
Philippa Middleton, honorary research fellow^o
Jacek Mrukowicz, executive director^p
Dianne O’Connell, senior epidemiologist^q
*Andrew D Oxman, director^f
Bob Phillips, associate fellow^f
*Holger J Schünemann, associate professor^g,s
Tessa Tan-Torres Edejer, medical officer/scientist^t
Helena Varonen, associate editor^u
Gunn E Vist, researcher^f
John W Williams Jr, associate professor^r
Stephanie Zaza, project director^w
~Prof. Patrick M Bossuyt and Prof. Victor M. Montori

x) Vanderbilt University Medical Center, USA
a) Agency for Healthcare Research and Quality, USA
b) Children's National Medical Center, USA
c) Centers for Disease Control and Prevention, USA
d) University of Newcastle upon Tyne, UK
e) German Cochrane Centre, Germany
f) Norwegian Centre for Health Services, Norway
g) McMaster University, Canada
h) Scottish Intercollegiate Guidelines Network, UK
i) Fédération Nationale des Centres de Lutte Contre le Cancer, France
j) University of Newcastle, Australia
k) McMaster University, Canada
l) National Institute for Clinical Excellence, UK
m) Università di Modena e Reggio Emilia, Italy
n) Centro per la Valutazione della Efficacia della Assistenza Sanitaria, Italy
o) Australasian Cochrane Centre, Australia
p) Polish Institute for Evidence Based Medicine, Poland
q) The Cancer Council, Australia
r) Centre for Evidence-based Medicine, UK
s) University of Buffalo, USA
t) World Health Organisation, Switzerland
u) Finnish Medical Society Duodecim, Finland
v) Duke University Medical Center, USA
w) Centers for Disease Control and Prevention, USA
Which tests should be recommended, and with what strength?

Why grade recommendations?

A systematic and explicit approach to making judgments about the quality of evidence and the strength of recommendations can help to prevent errors, facilitate critical appraisal of these judgments, and can help to improve communication of this information.
<table>
<thead>
<tr>
<th>Strong recommendations</th>
<th>Weak recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>- strong methods</td>
<td>- weak methods</td>
</tr>
<tr>
<td>- accurate test impacts outcome</td>
<td>- imprecise estimate / small effect</td>
</tr>
<tr>
<td>- few downsides of testing strategy</td>
<td>- substantial downsides</td>
</tr>
<tr>
<td>- indicating a judgment that a majority of well informed people will make the same</td>
<td>- indicating a judgment that a majority of well informed people will make the same</td>
</tr>
<tr>
<td>choice</td>
<td>choice, but a substantial minority will not (significant uncertainty)</td>
</tr>
<tr>
<td>- (high confidence, low uncertainty)</td>
<td>- expect variability in clinician and patient actions</td>
</tr>
<tr>
<td>- expect non-variant clinician and patient behavior - most patients should receive</td>
<td>- clinical expertise important - focus on decision-making and implementation</td>
</tr>
<tr>
<td>the intervention</td>
<td>- patient values and preferences important - focus on determining values and</td>
</tr>
<tr>
<td>- diminished role for clinical expertise - focus on implementation &amp; barriers</td>
<td>preferences relative to decision</td>
</tr>
<tr>
<td>- focused role of patient values and preferences - emphasis on compliance and</td>
<td>- decision aids likely to be useful</td>
</tr>
<tr>
<td>barriers</td>
<td>- offering the intervention and helping patients make a decision</td>
</tr>
<tr>
<td>- could be used as a performance / quality indicator</td>
<td>- could be used a quality criterion</td>
</tr>
<tr>
<td>- decision aids not likely to be needed</td>
<td>- medical practice is expected to vary to some degree</td>
</tr>
<tr>
<td>- medical practice is expected to not vary much</td>
<td>- medical practice is expected to vary to some degree</td>
</tr>
</tbody>
</table>
Grading the Evidence: Evaluating Diagnostic Studies

- Evidence concepts
  - scientific results that approximate truth
  - size, accuracy, precision
  - reliability, reproducibility, appropriateness, bias
  - statistical descriptions
  - trade-offs, limiting factors, cost

- Grade components
  - **Quality** (Validity)
    - The quality of evidence indicates the extent to which one can be confident that an estimate of effect is correct.
  - **Strength** (Benefit/Risk - Results)
    - The strength of a recommendation indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm.

- Implementation and application
  - Will the results help me with my patient care? (Relevance & Prevalence)
## Grading evidence process

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe question / problem (PICO)</td>
<td></td>
</tr>
<tr>
<td>Establish critical / important outcomes to decision and relative</td>
<td></td>
</tr>
<tr>
<td>importance of outcomes</td>
<td></td>
</tr>
<tr>
<td>Systematic review</td>
<td></td>
</tr>
<tr>
<td>Evidence profile / quality for each outcome</td>
<td></td>
</tr>
<tr>
<td>Overall quality of evidence</td>
<td></td>
</tr>
<tr>
<td>Benefit/risk/harm/cost balance</td>
<td></td>
</tr>
<tr>
<td>Overall strength of recommendation - GRADE</td>
<td></td>
</tr>
<tr>
<td>Implementation and evaluation</td>
<td></td>
</tr>
</tbody>
</table>
Evaluating Diagnosis Studies: Validity (Quality)

- Are the results Valid? (grading quality)
  - Was there an independent, blind comparison with a reference standard? (gold standard)
  - Did the patient sample include an appropriate spectrum of the sort of patients to whom the diagnostic test will be applied in clinical practice?
  - Is there a standard method for doing the test? (reproducibility, reliability)
### Judgments about Evidence Quality

The quality of evidence indicates the extent to which one can be confident that an estimate of effect is correct.

<table>
<thead>
<tr>
<th>Quality Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
<td>Blinded testing of previously developed diagnostic criteria in series of consecutive patients (with universally applied reference &quot;gold&quot; standard) or a systematic review of these studies</td>
</tr>
<tr>
<td><strong>Moderate</strong></td>
<td>Development of diagnostic criteria on basis of consecutive patients (with universally applied reference &quot;gold&quot; standard) or a systematic review of these studies</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>Study of nonconsecutive patients (no consistently applied reference &quot;gold&quot; standard) or a systematic review of these studies</td>
</tr>
<tr>
<td><strong>Very Low</strong></td>
<td>Study with a poor reference standard / Case-control study / Expert opinion / Case Report</td>
</tr>
</tbody>
</table>

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group

www.gradeworkinggroup.org
Types of studies to evaluate a test or diagnostic strategy: outcome-based

- Example: RCT that explored a diagnostic strategy guided by the use of B-type natriuretic peptide (BNP)
- Designed to provide a more accurate diagnosis of heart failure - in patients presenting to the emergency department with acute dyspnea.
- The group randomized to receive BNP spent a shorter time in the hospital at lower cost, with no increased mortality or morbidity.
- When diagnostic intervention studies - ideally RCTs but also observational studies - comparing alternative diagnostic strategies with assessment of direct patient important outcomes are available, guideline recommendation panels can use the GRADE approach established for treatment questions.

Types of studies to evaluate a test or diagnostic strategy: accuracy-based

- Diagnostic accuracy is a surrogate outcome for what we are really interested in, which is patient important benefit and harm.

- Example: consistent evidence from well-designed studies of fewer false negative results with non-contrast helical CT than with IVP in the diagnosis of acute urolithiasis.

- However, the stones in the ureters “missed” by IVP are smaller, and hence are likely to pass more easily.

- Since randomized trials evaluating outcomes in patients treated for smaller stones are not available, the extent to which CT would reduce “missed” cases (false negatives) and have important health benefits - remains uncertain.


Judgments about Evidence Quality

The quality of evidence indicates the extent to which one can be confident that an estimate of effect is correct.

- **High**
- **Moderate**
- **Low**
- **Very Low**

**Moving Down**:
- serious limitations in study design or execution, sparse data: serious flaws can lower by one level, fatal flaws can lower by two levels
- consistency: important inconsistency can lower by one level
- directness of evidence: some uncertainty lower by one level, major uncertainty lower by two levels
- selection bias or reporting bias: strong evidence lower by 1 level
- imprecise evidence, wide CI can lower by one level

adapted from Gordon Guyatt
Judgments about Evidence Quality

The quality of evidence indicates the extent to which one can be confident that an estimate of effect is correct.

- **High**: Further research is very unlikely to change our confidence in the estimates of diagnostic value.
- **Moderate**: Further research is likely to have an important impact on our confidence in the estimate of diagnostic value and may change the estimate.
- **Low**: Further research is very likely to have an important impact on our confidence in the estimate of diagnostic value and is likely to change the estimate.
- **Very Low**: Any estimate of effect is very uncertain.

GRADE’s four categories of quality of evidence imply a gradient of confidence in estimates of the effect of a diagnostic test or strategy on patient-important outcomes.
## Evidence profile / quality for each outcome

### Quality assessment of diagnostic accuracy studies – example: should multislice spiral computed tomography versus conventional coronary angiography be used for diagnosis of coronary artery disease?¹

<table>
<thead>
<tr>
<th>All outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No of studies</strong></td>
</tr>
<tr>
<td>21 studies (1,570 patients)</td>
</tr>
</tbody>
</table>

¹ all patients were selected to undergo conventional coronary angiography and were, therefore, presenting with high probability of coronary artery disease (median prevalence in the included studies: 63.5%, Range 6.6 to 100%)

² there was significant heterogeneity of results (not investigated further by the authors of the review)

³ the possibility of reporting bias exists but it was not considered sufficient to downgrade
Evaluating Diagnosis Studies: Strength (Results)

- What are the Results? (grading strength)
  - What is the magnitude of benefit, and how reliable/precise are these results?
  - What are the magnitudes of risk, burden, and cost; and how reliable/precise are these results?
  - Do the benefits outweigh the risks/burdens/costs? Are there known trade-offs? Are there unknown possible trade-offs?

- Result Parameters: Accuracy, Likelihood Ratios, Confidence Intervals
  - Are likelihood ratios for the test results presented or data necessary for their calculation included?
  - Statistics for Pre-test Probability (prevalence), Likelihood Ratios, Sensitivity and Specificity, Predictive Values

- Are the results of the test useful?
Benefit/risk/harm/cost balance

Strength

Net benefits: The test intervention does more good than harm.

Trade-offs: There are important trade-offs between the benefits and harms.

Uncertain trade-offs: Not clear whether the test intervention does more good than harm.

No net benefits: The test intervention does not do more good than harm.

Net harms: The test intervention does more harm than good.

Benefits

Absolute Benefit Increase X Utility

Desirable effects: health benefits, less burden, savings

Uncertain Trade-Offs

Risks & Harms

Absolute Risk Increase X Utility (Value Factor)

Undesirable effects: harms, more burden, costs

Net Benefits

Net Harms

Trade-Offs

Uncertain Trade-Offs

The strength of a recommendation indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm.
• Tests or test strategies that result in patients moving below the test threshold or above the treatment threshold (given the treatment exists) will often lead to strong recommendations despite the false negative and false positive test results.

• On the other hand, test or strategies that will only marginally change the probability of disease and require further testing will usually lead to weak recommendations.

• The test properties that best coincide with results that move the patient’s probability significantly up or down are the Likelihood Ratios.
Judgments about Overall Grade of Recommendations

- Judgments about the strength of a recommendation (Strong or Weak, For or Against) require consideration of:
  - all critical outcomes
    - (must be critical to care, critical to decision, not just important)
  - the quality of the evidence
    - the lowest quality of evidence for any critical outcome should provide the basis for grading
  - the balance between benefits and harms
    - if information on harm is critical, it should be included even if uncertainty exists
  - translation of the evidence into specific circumstances
    - evidence is global, application is local
  - the certainty of the baseline risk

- Also important to consider costs (resource utilization) prior to making a recommendation
Conclusions

• The GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic guidelines provides comprehensive and transparent methodology for developing these recommendations.

• The GRADE Working Group presents an overview of the approach, already established for grading treatment recommendations.

• Publication for Diagnostic GRADE in BMJ is pending.

• Extensive application to diagnostic guidelines is likely to refine the approach.

• The basic methodology & considerations that follow from recognizing test results as surrogate markers are unlikely to change.


