Guideline and policy development using the GRADE approach

Karen R Steingart, MD, MPH
karenst@uw.edu
Chennai, 15 December 2010

Acknowledgments

• This presentation is based on a workshop, “Teaching evidence assimilation for collaborative healthcare” the New York Academy of Medicine, August 2010

• Slides are used by permission of Holger Schünemann
Overview

• Describe background about GRADE

• Discuss factors influencing the quality of evidence

• Discuss the process of moving from evidence to recommendations

• Describe the WHO guideline development process using the GRADE approach

“Evidence does not make decisions, people do”

(Clinical) state and circumstances

Population values and preferences

Expertise

Research evidence

Haynes. BMJ 2002;324:1350
**Which hierarchy? (1)**

Recommendation for use of oral anticoagulation in patients with atrial fibrillation and rheumatic mitral valve disease

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Recommendation</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Class I</td>
<td>AHA</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>ACCP</td>
</tr>
<tr>
<td>IV</td>
<td>C</td>
<td>SIGN</td>
</tr>
</tbody>
</table>

---

**Centers for Disease Control and Prevention (CDC)**

<table>
<thead>
<tr>
<th>Evidence of Effectiveness</th>
<th>Execution - Good or Fair</th>
<th>Design Suitability — Greatest, Moderate, or Least</th>
<th>Number of Studies</th>
<th>Consistent</th>
<th>Effect Sized</th>
<th>Expert Opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>Good</td>
<td>Greatest</td>
<td>At Least 2</td>
<td>Yes</td>
<td>Sufficient</td>
<td>Not Used</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>Greatest or Moderate</td>
<td>At Least 5</td>
<td>Yes</td>
<td>Sufficient</td>
<td>Not Used</td>
</tr>
<tr>
<td></td>
<td>Good or Fair</td>
<td>Greatest</td>
<td>At Least 5</td>
<td>Yes</td>
<td>Sufficient</td>
<td>Not Used</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meet Design, Execution, Number, and Consistency Criteria for Sufficient But Not Strong Evidence</td>
<td>Large</td>
<td></td>
<td></td>
<td>Not Used</td>
</tr>
<tr>
<td>Sufficient</td>
<td>Good</td>
<td>Greatest</td>
<td>1</td>
<td>Not Applicable</td>
<td>Sufficient</td>
<td>Not Used</td>
</tr>
<tr>
<td></td>
<td>Good or Fair</td>
<td>Greatest or Moderate</td>
<td>At Least 3</td>
<td>Yes</td>
<td>Sufficient</td>
<td>Not Used</td>
</tr>
<tr>
<td></td>
<td>Good or Fair</td>
<td>Greatest, Moderate, or Least</td>
<td>At Least 5</td>
<td>Yes</td>
<td>Sufficient</td>
<td>Not Used</td>
</tr>
<tr>
<td>Expert Opinion</td>
<td>Varies</td>
<td>Varies</td>
<td>Varies</td>
<td>Varies</td>
<td>Sufficient</td>
<td>Supports a Recommendation</td>
</tr>
<tr>
<td>Insufficient</td>
<td>A. Insufficient Designs or Execution</td>
<td>B. Too Few Studies</td>
<td>C. Inconsistent</td>
<td>D. Small</td>
<td>E. Not Used</td>
<td></td>
</tr>
</tbody>
</table>
Which hierarchy? (2)

**STUDY DESIGN**
- Randomized Controlled Trials
- Cohort Studies and Case Control Studies
- Case Reports and Case Series, Non-systematic observations
- Expert Opinion

**BIAS**
United States Parachute Association reported 821 injuries and 18 deaths out of 2.2 million jumps in 2007; relative risk reduction > 99.9% (1/100,000)

Simple hierarchies are (too) simplistic

**STUDY DESIGN**
- Randomized Controlled Trials
- Cohort Studies and Case Control Studies
- Case Reports and Case Series, Non-systematic observations
- Expert opinion

**BIAS**

Expert Opinion
The Grading of Recommendations Assessment, Development and Evaluation (GRADE)

• Aim: to develop a common, transparent and sensible system for grading the quality of evidence and the strength of recommendations
• International group of guideline developers, methodologists & clinicians from around the world (>100 contributors) since 2000
• International group: ACCP, AHRQ, Australian NMRC, BMJ Clinical Evidence, CC, CDC, McMaster, NICE, Oxford CEBM, SIGN, UpToDate, USPSTF, WHO


GRADE Uptake

- World Health Organization
- Allergic Rhinitis in Asthma Guidelines (ARIA)
- American Thoracic Society
- American College of Physicians
- European Respiratory Society
- European Society of Thoracic Surgeons
- British Medical Journal
- Infectious Disease Society of America
- American College of Chest Physicians
- UpToDate®
- National Institutes of Health and Clinical Excellence (NICE)
- Scottish Intercollegiate Guideline Network (SIGN)
- Cochrane Collaboration
- Infectious Disease Society of America
- Clinical Evidence
- Agency for Health Care Research and Quality (AHRQ)
- Partner of GIN
- Over 40 major organizations
Types of questions

Background Questions
• Definition: What is latent TB infection?
• Mechanism: How does an IGRA work?

Foreground Questions
• Benefit > harm: Does the use of IGRAs improve the identification of HIV-infected individuals who could benefit from treatment of LTBI?

Framing a foreground question
• Population: Individuals with/suspected of LTBI
• Intervention: IGRA
• Comparison: No test/other IGRA, TST
• Outcomes: Survival, mortality, development of TB disease, hospitalizations, resource use, adverse outcomes, antimicrobial resistance

Schunemann, et al., The Lancet ID, 2007
GRADE rating of outcomes

• GRADE rates the quality of evidence for each outcome separately
  – The type of evidence may be different for different outcomes

• GRADE considers desirable and undesirable outcomes and rates their relative importance

Outcomes may be desirable or undesirable

• Desirable outcomes
  - Decreased mortality
  - Reduced duration of disease
  - Reduced resource expenditure

• Undesirable outcomes
  - Adverse events
  - The development of resistance
  - Costs of treatment

• Every decision comes with desirable/undesirable consequences
• Developing recommendations must include a consideration of desirable and undesirable outcomes
What is quality?

• “In the context of making recommendations, the quality of evidence reflects the extent to which our confidence in an estimate of the effect is adequate to support a particular recommendation.” Gordon Guyatt BMJ 2008
Definition of grades of evidence

- /High: Further research is very unlikely to change confidence in the estimate of effect
- /Moderate: Further research is likely to have an important impact on confidence in the estimate of effect and may change the estimate
- /Low: Further research is very likely to have an important impact on confidence in the estimate of effect and is likely to change the estimate
- /Very low: Any estimate of effect is very uncertain

There always is evidence

“When there is a question there is evidence”

Better research ⇒ greater confidence in the evidence and decisions
Controlled trial of fasting and one-year vegetarian diet in rheumatoid arthritis. The Lancet; 338: 899-902

Authors’ conclusions: Short term beneficial effects were found for fasting for 7 to 10 days followed by a vegetarian diet when compared to ordinary diet

GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis, 2008

Forest plot of sensitivity (a) and specificity (b) estimates for rifampicin resistance

Authors’ conclusions: GenoType MTBDR assays demonstrate excellent accuracy for rifampicin resistance
- What information do you think would increase or decrease your confidence in these results?
- What information do you think would indicate that more research is or is not necessary?
Determinants of quality for diagnostic questions

• RCTs and observational studies: start high if direct ☐☐☐☐
• 5 factors can lower quality
  1. Limitations in detailed design and execution (risk of bias criteria)
  2. Inconsistency (or heterogeneity)
  3. Indirectness (PICO and applicability)
  4. Imprecision (number of events and confidence intervals)
  5. Publication bias
• 3 factors can increase quality
  1. Large magnitude of effect
  2. Plausible residual confounding may be working to reduce the demonstrated effect or increase the effect if no effect was observed
  3. Dose-response gradient

1. Design and Execution/Risk of Bias (QUADAS)

Examples:
• Was an unselected sample of patients enrolled? (consecutive with suspected disease)

• Were the index test results interpreted without knowledge of the results of the reference standard?

• Did all patients receive the same reference standard?
Methodological quality summary: review authors’ judgments about each methodological quality item for each included study, created in RevMan http://ims.cochrane.org/revman

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Oraiey 1992a</td>
<td>✔</td>
<td>✔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>✔</td>
</tr>
<tr>
<td>Al-Oraiey 1992b</td>
<td>✔</td>
<td>✔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>✔</td>
</tr>
<tr>
<td>Al-Namni 2009a</td>
<td>✗</td>
<td>✔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>✔</td>
</tr>
<tr>
<td>Al-Namni 2009b</td>
<td>✔</td>
<td>✔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>✔</td>
</tr>
<tr>
<td>Atay 1993a</td>
<td>✔</td>
<td>✔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>❔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Who believes the risk of bias is of concern?

- Yes
- No
- Don’t know or undecided
Would you downgrade for risk of bias?

- No, there are no serious limitations
- Yes, there are serious limitations
- Yes, there are very serious limitations

2. Inconsistency of results (Heterogeneity)

- if inconsistency, look for explanation
  - patients, intervention, comparator, outcome
- if unexplained inconsistency lower quality
Forest plot sensitivity, serological tests, smear-negative pulmonary TB patients

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfano 1996c</td>
<td>0.64 (0.46 - 0.72)</td>
</tr>
<tr>
<td>Alfano 1996d</td>
<td>0.84 (0.46 - 0.73)</td>
</tr>
<tr>
<td>Alfaro 1997a</td>
<td>0.68 (0.44 - 0.92)</td>
</tr>
<tr>
<td>Andrussov-1995b</td>
<td>0.63 (0.30 - 0.73)</td>
</tr>
<tr>
<td>Chandrasekaran 1996</td>
<td>0.55 (0.42 - 0.67)</td>
</tr>
<tr>
<td>Imaiz 2004d</td>
<td>0.49 (0.32 - 0.66)</td>
</tr>
<tr>
<td>Imaiz 2004b</td>
<td>0.32 (0.14 - 0.49)</td>
</tr>
<tr>
<td>Imaiz 2004c</td>
<td>0.34 (0.20 - 0.50)</td>
</tr>
<tr>
<td>Imaiz 2004d</td>
<td>0.83 (0.47 - 0.79)</td>
</tr>
<tr>
<td>Imaiz 2004e</td>
<td>0.59 (0.42 - 0.74)</td>
</tr>
<tr>
<td>Imaiz 2004f</td>
<td>0.84 (0.53 - 0.95)</td>
</tr>
<tr>
<td>Imaiz 2004g</td>
<td>0.68 (0.42 - 0.78)</td>
</tr>
<tr>
<td>Imaiz 2004h</td>
<td>0.29 (0.10 - 0.49)</td>
</tr>
<tr>
<td>Imaiz 2004i</td>
<td>0.59 (0.40 - 0.79)</td>
</tr>
<tr>
<td>Imaiz 2004j</td>
<td>0.51 (0.35 - 0.67)</td>
</tr>
<tr>
<td>Imaiz 2004k</td>
<td>0.49 (0.33 - 0.66)</td>
</tr>
<tr>
<td>Imaiz 2004l</td>
<td>0.66 (0.49 - 0.83)</td>
</tr>
<tr>
<td>Imaiz 2004m</td>
<td>0.61 (0.45 - 0.78)</td>
</tr>
<tr>
<td>Imaiz 2004n</td>
<td>0.76 (0.60 - 0.88)</td>
</tr>
<tr>
<td>Lufi 1995</td>
<td>0.71 (0.59 - 0.82)</td>
</tr>
<tr>
<td>Makens 2002b</td>
<td>0.69 (0.55 - 0.81)</td>
</tr>
<tr>
<td>Makens 2002c</td>
<td>0.90 (0.80 - 1.00)</td>
</tr>
<tr>
<td>Nakvaspadt et al. 2002b</td>
<td>0.46 (0.27 - 0.65)</td>
</tr>
<tr>
<td>Okuda 2004b</td>
<td>0.73 (0.52 - 0.95)</td>
</tr>
<tr>
<td>Okuda 2004c</td>
<td>0.68 (0.37 - 0.77)</td>
</tr>
<tr>
<td>Okuda 2004d</td>
<td>0.68 (0.38 - 0.92)</td>
</tr>
<tr>
<td>Okuda 2004f</td>
<td>0.76 (0.64 - 0.88)</td>
</tr>
<tr>
<td>Vito 2001b</td>
<td>0.35 (0.25 - 0.46)</td>
</tr>
</tbody>
</table>

Pooled Sensitivity = 0.65 (0.55 to 0.76)
Chi-square = 56.53, df = 27 (p < 0.0000)
Inconsistency (I²) = 72.5%

3. Indirectness

- The extent to which the study’s patients, interventions, and outcomes are similar to those in practice.

Examples
- differences in populations (study involves adults, can you generalize to children?)
- Differences in settings (interested in low income, but all data come from high income)
- No head to head comparisons
Test accuracy is a *surrogate* for patient important outcomes

- When clinicians think about diagnostic tests, they focus on test accuracy, e.g., sensitivity/specificity

- The underlying assumption is that knowing whether a target condition is present or absent will result in superior patient management and improved outcomes....But does it?

4. Imprecision

- Reliability of an estimate of effect
- Best described by the width of the 95% CI
- Precision is influenced by the sample size of the study
5. Publication Bias

- Should always be suspected
  - Only small “positive” trials
  - For profit interest
  - Various methods to evaluate for systematic reviews of interventions, no agreed upon method for diagnostic reviews

GRADE evidence profile

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>No. studies (Participants)</th>
<th>Study Design</th>
<th>Limitations</th>
<th>Indicators</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Publication Bias</th>
<th>Quality</th>
<th>Effect per 1000</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives</td>
<td>17 (2016)</td>
<td>Cross-sectional and case-control</td>
<td>Very</td>
<td>Sensitivity (2)</td>
<td>No</td>
<td>No</td>
<td>Likely</td>
<td>Very Low</td>
<td>0/1000</td>
<td>Critical</td>
</tr>
<tr>
<td>True Negatives</td>
<td>17 (2016)</td>
<td>Cross-sectional and case-control</td>
<td>Very</td>
<td>Sensitivity (2)</td>
<td>No</td>
<td>No</td>
<td>Likely</td>
<td>Very Low</td>
<td>0/1000</td>
<td>Critical</td>
</tr>
<tr>
<td>False Positives</td>
<td>17 (2016)</td>
<td>Cross-sectional and case-control</td>
<td>Very</td>
<td>Sensitivity (2)</td>
<td>No</td>
<td>No</td>
<td>Likely</td>
<td>Very Low</td>
<td>0/1000</td>
<td>Critical</td>
</tr>
<tr>
<td>False Negatives</td>
<td>17 (2016)</td>
<td>Cross-sectional and case-control</td>
<td>Very</td>
<td>Sensitivity (2)</td>
<td>No</td>
<td>No</td>
<td>Likely</td>
<td>Very Low</td>
<td>0/1000</td>
<td>Critical</td>
</tr>
</tbody>
</table>

Footnotes

Steingart et al, submitted manuscript
Moving From Evidence to Recommendations

<table>
<thead>
<tr>
<th></th>
<th>Cross-sectional and case-control</th>
<th>Very Serious&lt;sup&gt;A1&lt;/sup&gt; (-2)</th>
<th>No Serious Indirectness&lt;sup&gt;A2&lt;/sup&gt;</th>
<th>Very Serious&lt;sup&gt;A3&lt;/sup&gt; (-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives</td>
<td>67 (8318)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Negatives</td>
<td>67 (8318)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on sensitivity median = 64%, specificity median = 91%

<sup>A1</sup> Majority of studies lacked a representative patient spectrum and were not blinded.

<sup>A2</sup> Although diagnostic accuracy is considered a surrogate for patient-important outcomes, we did not.

<sup>A3</sup> There was considerable heterogeneity in study results.

<sup>A4</sup> We did not pool accuracy estimates. The 95% CIs were wide for many individual studies. We did not

2 points for inconsistency.
Strength of recommendation

• “The strength of a recommendation reflects the extent to which we can, across the range of patients for whom the recommendations are intended, be confident that desirable effects of a management strategy outweigh undesirable effects”

• Recommendations may be strong or weak/conditional

Developing recommendations

The figure describes the balance between important benefits and downsides relate to a recommendation. The process begins by evaluating whether desirable effects outweigh undesirable effects or vice versa. Moving on to making a recommendation requires a decision. If the balance is clear, a strong recommendation for or against an action follows (<= and => denote a clear balance). If the balance is not clear, a weak recommendation for or against an action follows (?< and ?>> denote a balance that is not clear). Widely differing values (the importance or preference patients assign to a certain health state) can also lead to a less clear balance of benefits versus downsides.
Determinants of strength of recommendation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance between desirable and undesirable effects</td>
<td>The larger the difference between desirable and undesirable effects, the higher the likelihood that a strong recommendation is warranted</td>
</tr>
<tr>
<td>Quality of evidence</td>
<td>The stronger the quality of evidence, the higher the likelihood that a strong recommendation is warranted</td>
</tr>
<tr>
<td>Values and preferences</td>
<td>The more values and preferences vary, the higher the likelihood that a weak recommendation is warranted</td>
</tr>
<tr>
<td>Costs (resource allocation)</td>
<td>The higher the costs, that is the greater the resources consumed, the lower the likelihood that a strong recommendation is warranted</td>
</tr>
</tbody>
</table>

Implications of a strong recommendation

• Patients: Most people in this situation would want the recommended course of action and only a small proportion would not
• Clinicians: Most patients should receive the recommended course of action
• Policy makers: The recommendation can be adapted as a policy in most situations
Implications of a conditional/weak recommendation

- Patients: The majority of people in this situation would want the recommended course of action, but many would not.

- Clinicians: Be more prepared to help patients to make a decision that is consistent with their own values/decision aids and shared decision making.

- Policy makers: There is a need for substantial debate and involvement of stakeholders.

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Patients with HIV and drug resistant TB requiring second line drugs, the expert panel recommends/suggests to [not] administer ART (7 recommendation, 4 quality evidence).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>HIV positive individuals with drug resistant TB requiring second line drugs</td>
</tr>
<tr>
<td>Intervention</td>
<td>ART for during TB treatment versus ART salvage</td>
</tr>
<tr>
<td>Factor</td>
<td>Decision</td>
</tr>
<tr>
<td>High or moderate quality evidence ([7 high quality evidence])</td>
<td>Yes</td>
</tr>
<tr>
<td>Certainty about the balance of benefits versus harms and burdens ([3 more certainty])</td>
<td>Yes</td>
</tr>
<tr>
<td>Certainty/similarity in values ([3 more certainty])</td>
<td>Yes</td>
</tr>
<tr>
<td>Resource implications ([2 more resources required for concordant ART use])</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Matt Arentz et al, unpublished
What is a WHO guideline?

- “A WHO guideline is any document containing recommendations about health interventions, whether they are clinical, public health or policy.”

- World Health Organization Handbook for guideline development, March 2008
In Summary

• Guidelines should be based on the best available evidence to be evidence-based
• GRADE combines health research methodology with a structured approach to improve communication
• Criteria for evidence assessment across questions and outcomes
• Criteria for moving from evidence to recommendations
• Transparent, systematic
  four categories of quality of evidence
  two grades for strength of recommendations
• Transparency in decision making and judgments is key