Guideline development in TB diagnostics

Karen R Steingart, MD, MPH

July 10, 2014
karen.steingart@gmail.com
Conflicts of interest

• Editor Cochrane Infectious Diseases Group

• Editor Cochrane Diagnostic Test Accuracy Working Group

• Member GRADE Working Group

• No financial interests to declare
Overview

- Describe the World Health Organization Global TB Programme guideline development process for TB diagnostics

- Describe the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach for developing guidelines

- Discuss challenges in developing guidelines
What are guidelines?

• Guidelines are recommendations intended to assist providers and recipients of health care and other stakeholders to make informed decisions. Recommendations may relate to clinical interventions, public health activities, or government policies. WHO 2003, 2007
Use of evidence in WHO recommendations

Andrew D Oxman, John N Lavis, Atle Fretheim

Summary

Background WHO regulations, dating back to 1951, emphasise the role of expert opinion in the development of recommendations. However, the organisation's guidelines, approved in 2003, emphasise the use of systematic reviews for evidence of effects, processes that allow for the explicit incorporation of other types of information (including values), and evidence-informed dissemination and implementation strategies. We examined the use of evidence, particularly evidence of effects, in recommendations developed by WHO departments.

Methods We interviewed department directors (or their delegates) at WHO headquarters in Geneva, Switzerland, and reviewed a sample of the recommendation-containing reports that were discussed in the interviews (as well as related background documentation). Two individuals independently analysed the interviews and reviewed key features of the reports and background documentation.

Findings Systematic reviews and concise summaries of findings are rarely used for developing recommendations. Instead, processes usually rely heavily on experts in a particular specialty, rather than representatives of those who will have to live with the recommendations or on experts in particular methodological areas.

Interpretation Progress in the development, adaptation, dissemination, and implementation of recommendations for member states will need leadership, the resources necessary for WHO to undertake these processes in a transparent and defensible way, and close attention to the current and emerging research literature related to these processes.
WHO appraisal of key criteria for guidelines

- Transparency
- End-users
- Conflict of interest
- Systematic reviews
- Expert opinion
- Resources

“I would have liked to have had more evidence to base recommendations on.”

Oxman et al. Lancet 2007
In response to concerns about the quality of WHO guidelines, and following up on recommendations by The Advisory Committee on Health Research (ACHR) and resolution EB120.R15 of the 120th Session of the Executive Board, this note announces the establishment of a WHO Guidelines Review Committee (GRC). The GRC will develop and implement standards and procedures for guideline development that ensure that WHO guidelines are consistent with internationally accepted best practice, including appropriate use of evidence.
WHO standards and procedures for guidelines

- Plan and scope out questions
- Form guideline development group and manage declarations of interest
- Formulate PICO questions
- Retrieve evidence/systematic reviews
- Assess quality of evidence and develop recommendations (GRADE)
- Writing and external review
- Disseminate, implement, evaluate
- Update
The WHO Global TB Programme
Core Functions

• Provide global leadership on matters critical to TB

• **Develop evidence-based policies, strategies and standards for TB prevention, care and control, and monitor their implementation**

• Jointly with WHO regional and country offices, provide technical support to Member States, catalyze change, and build sustainable capacity

• Monitor the global TB situation, and measure progress in TB care, control, and financing

• Shape the TB research agenda and stimulate the generation, translation and dissemination of valuable knowledge

• Facilitate and engage in partnerships for TB action
WHO TB diagnostics policy formulation process

1. Identifying the need for policy change
   - WHO strategic monitoring of country needs
   - Partners (researchers, industry, etc)
   - Body of evidence available

2. Reviewing the evidence
   - Commissioning of systematic reviews
   - QUADAS or other diagnostic accuracy tool
   - Meta-analyses (where feasible)

3. Convening an Expert Group
   - Experts, methodologists, end-users
   - Guidelines Review Committee
   - GRADE process for evidence synthesis

4. Assessing policy proposal and recommendations
   - Strategic and Technical Advisory Group
   - Endorsement/revision/addition
   - Advise to WHO to proceed/not with policy
   - Evolution to ‘negative’ policy

5. Formulating and disseminating policy
   - Guidelines Review Committee
   - Dissemination to Member States
   - Promotion with stakeholders & funders
   - Phased implementation & scale-up plan

who.int/tb/advisory_bodies/research_to_policy/en/index.html
WHO policy statements on TB diagnostics

- Definition of a new sputum smear-positive TB case (2007)
- Reduction of number of smears for diagnosis of pulmonary TB (2007)
- Automated liquid culture and drug susceptibility testing (2007)
- Molecular line probe assays for rapid screening of patients at risk of MDR-TB (2008)
- Drug susceptibility testing second-line drugs (2008)
- Same-day diagnosis of tuberculosis by microscopy (2010)
- Fluorescence LED microscopy (2010)
- Non-commercial drug susceptibility testing methods for screening patients at risk for MDR-TB (2010)
- Commercial serodiagnostic tests (2011)
- Interferon-gamma release assays in low- and middle-income countries (2011)
- Xpert MTB/RIF (2011)
- Xpert MTB/RIF update (2013)
Convene a knowledgeable, multidisciplinary panel of experts and representatives from affected groups.
The Vexing Problem of Guidelines and Conflict of Interest: A Potential Solution

Gordon Guyatt, MD, MSc; Elie A. Akl, MD, PhD; Jack Hirsch, MD; Clive Keation, MD, PhD; Mark Crowther, MD; David Gutterman, MD; Sandra Zelman Lewis, PhD; Ian Nathanson, MD; Roman Jaeschke, MD, MSc; and Holger Schünemann, MD, PhD

Health Research Policy and Systems

Review
Improving the use of research evidence in guideline development:
4. Managing conflicts of interests
Elizabeth A Boyd*1 and Lisa A Bero2

ICMJE Form for Disclosure of Potential Conflicts of Interest

Instructions
The purpose of this form is to provide readers of your manuscript with information about your other interests that could influence how they receive and understand your work. The form is designed to be completed electronically and stored electronically. It contains programming that allows appropriate data display. Each author should submit a separate form and is responsible for the accuracy and completeness of the submitted information. The form is in four parts.

1. Identifying Information.
Enter your full name. If you are NOT the corresponding author please check the box "no" and a space to enter the name of the corresponding author in the space that appears. Provide the requested manuscript information. Double-check the manuscript number and enter it.
Conflict of interest

- **Financial** – receipt of income or support that is related to, or could be affected by, the outcome of the WHO meeting or activity in which they are involved.

- **Intellectual** - academic activities that create the potential for an attachment to a specific point of view that could unduly affect an individual’s judgment about a specific recommendation.

- **Examples**
  - Grants, fellowships, honoraria
  - Participation in research
  - Commentary directly related to the recommendation
  - Consultancies
The defining feature of GRADE is separating judgements about the confidence in estimates (quality of evidence) from judgements about the strength of recommendations.
The GRADE approach

- Developed by a broad group of international guideline developers
- Clear separation between judging confidence in the effect estimates and strength of recommendations
- Explicit evaluation of the importance of outcomes
- Comprehensive criteria for downgrading and upgrading quality of evidence ratings
- Structured process for moving from evidence to recommendations
- Explicit acknowledgment of values and preferences
- Clear interpretation of strong versus weak recommendations for clinicians, patients, and policy makers

GRADE is not a system for performing systematic reviews
Organizations using GRADE

- World Health Organization
- Advisory Committee on Immunization Practices
- American Thoracic Society
- American College of Physicians
- European Respiratory Society
- British Medical Journal
- American College of Chest Physicians
- UpToDate®
- Agency for Health Care Research and Quality (AHRQ)
- National Institutes of Health and Clinical Excellence (NICE)
- Scottish Intercollegiate Guideline Network (SIGN)
- The Cochrane Collaboration
- Infectious Disease Society of America
- Canadian Task Force on Preventive Health Care Clinical Evidence
- Partner of Guidelines International Network
- Over 60 major organizations
Evidence-based healthcare decisions

(Clinical) state and circumstances

Population values and preferences

Research evidence

Expertise

Haynes et al. 2002
Hierarchy of evidence

STUDY DESIGN
- Randomized Controlled Trials
- Cohort, Cross-Sectional, and Case-Control Studies
- Case Reports and Case Series, Non-systematic observations
- Expert opinion
Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

Gordon C S Smith, Jill P Pell

Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials
What this study adds

No randomised controlled trials of parachute use have been undertaken.

The basis for parachute use is purely observational, and its apparent efficacy could potentially be explained by a “healthy cohort” effect.

Individuals who insist that all interventions need to be validated by a randomised controlled trial need to come down to earth with a bump.
Parachute use and risk of death

US Parachute Association reported 24 deaths out of 3.2 million jumps in 2013

0.0075 fatalities per 1000 jumps (one fatality per 133,333 skydives)

Quality of evidence, consider
- magnitude
- confidence

Quality of evidence using the GRADE approach

Magnitude and confidence
Simple hierarchies are (too) simplistic

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

**BIAS**
Expert opinion to interpret evidence
Guideline development

**Formulate recommendations:**
- For or against (direction)
- Strong or weak (strength)

*By considering:*
- Quality of evidence
- Balance benefits/harms
- Values and preferences

**Revise if necessary by considering:**
- Resource use (cost)

**Rate overall quality of evidence across outcomes based on lowest quality of critical outcomes**

- “We recommend using…”
- “We suggest using…”
- “We recommend against using…”
- “We suggest against using…”
(PICO) PPPIRTR clinical questions for diagnostic tests

- **Participants**, **Presentation**, **Prior tests**
- **Index test(s)**
- **Role** of testing (triage, replacement, add-on)
- **Target condition**
- **Reference standard**
Asking a sensible clinical question

- Should TEST A vs. TEST B be used in SOME PATIENTS/POPULATION?
- Should TEST A vs. TEST B be used for SOME PURPOSE?

In patients suspected of TB meningitis (patients), should Xpert MTB/RIF (intervention) be used as a replacement for usual practice (comparison) for the diagnosis of TB in low- and middle-income countries?
Selecting outcomes important to patients

• **Favorable**
  - decreased mortality
  - decreased hospital stay
  - decreased resource expenditure

• **Unfavorable**
  - adverse events
  - increased drug resistance
Relative importance of outcomes

- Decision makers (and guideline authors) need to consider the relative importance of outcomes when balancing these outcomes to make a recommendation.
- Relative importance vary across populations.
- Relative importance may vary across patient groups within the same population.
- When considered critical - evaluate.
### Test accuracy outcomes

<table>
<thead>
<tr>
<th>Disease (Reference standard)</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index test</strong></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td><strong>True Positive</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>False Negative</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- **Sensitivity** = $\frac{TP}{TP + FN}$
- **Specificity** = $\frac{TN}{TN + FP}$

**True Positive (TP)**
**False Positive (FP)**
**True Negative (TN)**
**False Negative (FN)**
**TP + FP**
**FN + TN**
**TP + FP + FN + TN**
## Patient and public health outcomes

<table>
<thead>
<tr>
<th>Index test +</th>
<th>TB</th>
<th>Not TB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FP</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index test -</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TN</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Patient
1. Drug toxicities/interactions
2. Anxiety/stigma
3. Missed true diagnosis

### Public Health
4. Misallocated meds/labs/x-rays
5. Misallocated staff time

*Slide curtesy Lucian Davis*
Quality of evidence

In the context of making recommendations for guidelines, quality reflects our confidence that the effect estimates are adequate to support a particular recommendation.
GRADE specifies four categories for the quality of a body of evidence

<table>
<thead>
<tr>
<th>Quality level</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>High ⊕⊕⊕⊕⊕</td>
<td>We are very confident that the true effect lies close to that of the estimate of the effect</td>
</tr>
<tr>
<td>Moderate ⊕⊕⊕</td>
<td>We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different</td>
</tr>
<tr>
<td>Low ⊕⊕⊕⊕</td>
<td>Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect</td>
</tr>
<tr>
<td>Very low ⊕⊕⊕⊕⊕</td>
<td>We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect</td>
</tr>
</tbody>
</table>

H. Balshem et al. Journal of Clinical Epidemiology 2010
What information about studies in a systematic review may affect our confidence that the estimate of an effect is correct?

• Were randomized controlled trials or high quality cross-sectional studies included?
• How many studies were pooled to get this estimate?
• How many patients did they include?
• How wide were the CIs around the effect estimate?
• Did the studies have important limitations, such as lack of blinding?
Domains for downgrading

5 factors can lower quality

1. Limitations (in study design and execution)
2. Indirectness (the question being addressed differs from the available evidence)
3. Inconsistency (heterogeneity)
4. Imprecision (wide confidence intervals)
5. Publication bias (difficult to appraise in systematic reviews of diagnostic test accuracy)
Domains for upgrading

3 factors can increase quality

1. Magnitude of the effect

2. Plausible confounding can increase confidence in estimated effects

3. Dose-response gradient
1. Limitations in study design and execution

- Cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard are considered high quality and can move to moderate, low or very low depending on other factors.
QUADAS-2 criteria (risk of bias)

- Could the selection of patients have introduced bias?
- Could the conduct or interpretation of the index test have introduced bias?
- Could the reference standard, its conduct, or its interpretation have introduced bias?
- Could the patient flow have introduced bias?
2. Indirectness

Do we have direct evidence? The quality of evidence may be lowered if...

- differences exist between populations studied and those for whom the recommendation is intended
- differences exist in expertise of people applying the tests in the studies compared to those applying the tests in usual practice settings
- tests being compared are evaluated in different studies, not directly compared in the same study
- effect of the test on patient outcomes is unavailable; sensitivity/specificity is a proxy for patient outcomes
When sensitivity/specificity is a proxy for patient outcomes

A good way to assess a diagnostic test or strategy would be a test-treat RCT: allocate patients to experimental or control diagnostic strategies and measure patient-important outcomes (mortality, morbidity, symptoms, quality of life and resource use)

- When sensitivity/specificity is a proxy, the panel will need to infer about the consequences of falsely identifying patients as having or not having the disease
- If a test fails to improve patient outcomes there is no reason to use it, whatever its accuracy
3. Inconsistency

The quality of evidence can be lowered if...

- there is unexplained variability in the sensitivity or specificity estimates among studies (heterogeneity)
- there is little overlap in the confidence intervals of the accuracy estimates

*If study results are in the same direction with overlapping confidence intervals, inconsistency is unlikely*
Inconsistency?

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

Inconsistency?


<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alifano 1998b</td>
<td>31</td>
<td>3</td>
<td>11</td>
<td>41</td>
<td>0.74 [0.58, 0.86]</td>
<td>0.93 [0.81, 0.99]</td>
</tr>
<tr>
<td>Banerjee 2003a</td>
<td>13</td>
<td>13</td>
<td>17</td>
<td>19</td>
<td>0.43 [0.25, 0.63]</td>
<td>0.59 [0.41, 0.76]</td>
</tr>
<tr>
<td>Caminero 1993</td>
<td>16</td>
<td>0</td>
<td>14</td>
<td>48</td>
<td>0.53 [0.34, 0.72]</td>
<td>1.00 [0.93, 1.00]</td>
</tr>
<tr>
<td>Caminero 1994</td>
<td>18</td>
<td>2</td>
<td>38</td>
<td>29</td>
<td>0.32 [0.20, 0.46]</td>
<td>0.94 [0.79, 0.99]</td>
</tr>
<tr>
<td>Gevaudan 1992b</td>
<td>26</td>
<td>47</td>
<td>0</td>
<td>147</td>
<td>1.00 [0.87, 1.00]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevaudan 1992f</td>
<td>53</td>
<td>47</td>
<td>3</td>
<td>147</td>
<td>0.95 [0.85, 0.99]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevaudan 1992j</td>
<td>25</td>
<td>47</td>
<td>0</td>
<td>147</td>
<td>1.00 [0.86, 1.00]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Gevaudan 1992n</td>
<td>34</td>
<td>47</td>
<td>0</td>
<td>147</td>
<td>1.00 [0.90, 1.00]</td>
<td>0.76 [0.69, 0.82]</td>
</tr>
<tr>
<td>Kunter 2003a</td>
<td>23</td>
<td>5</td>
<td>65</td>
<td>32</td>
<td>0.26 [0.17, 0.37]</td>
<td>0.86 [0.71, 0.95]</td>
</tr>
<tr>
<td>Luh 1996</td>
<td>28</td>
<td>17</td>
<td>30</td>
<td>207</td>
<td>0.48 [0.35, 0.62]</td>
<td>0.92 [0.88, 0.96]</td>
</tr>
</tbody>
</table>
4. Imprecision

- Results may be imprecise when studies include relatively few patients and few patients with disease, and hence have wide confidence intervals around sensitivity or specificity estimates.

MTBDRsI detection of XDR-TB, indirect testing
- Pooled sensitivity 70.9% (95% CI 42.9, 88.8)
- Pooled specificity 98.8% (95% CI 96.1, 99.6)

Theron et al, The Cochrane Library, submitted
5. Publication Bias

• Formal assessment of publication bias using methods such as funnel plots or regression tests is not recommended for diagnostic test accuracy studies.

• Difficult to be confident that publication bias is absent.
# Abdominal ultrasound for the diagnosis of pancreatic cancer

**Patients or population:** symptomatic patients in primary care with suspicion of pancreatic cancer  
**Setting:** mainly outpatients  
**New Test:** abdominal ultrasound  
**Reference Test:** endoscopic ultrasound with biopsy  
**Threshold:** Proven or probable pancreatic cancer

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-Positive</td>
<td>Prevalence 20 per 1000: Which is typically seen in otherwise healthy adults presenting with symptoms of jaundice, fatigue, pain of the abdomen, and dark urine.</td>
<td>13 per 1000 (9 to 15 per 1000)</td>
<td>282 per 1000 (220 to 339 per 1000)</td>
</tr>
<tr>
<td>False positives</td>
<td>Prevalence 400 per 1000: Which is typically seen in older adults presenting with symptoms of jaundice, fatigue and pain, with a family history of pancreatic cancer, history of chronic pancreatitis, who have diabetes, and are current or past smokers.</td>
<td>49 per 1000 (30 to 89 per 1000)</td>
<td>28 per 1000 (17 to 50 per 1000)</td>
</tr>
<tr>
<td>Test-Negative</td>
<td>Prevalence 20 per 1000: Which is typically seen in otherwise healthy adults presenting with symptoms of jaundice, fatigue, pain of the abdomen, and dark urine.</td>
<td>931 per 1000 (901 to 960 per 1000)</td>
<td>532 per 1000 (510 to 543 per 1000)</td>
</tr>
<tr>
<td>False negatives</td>
<td>7 per 1000 (4 to 10 per 1000)</td>
<td>158 per 1000 (101 to 220 per 1000)</td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval

Footnotes:  
\textsuperscript{1} A diagnostic test for pancreatic cancer needs to be less invasive than the current reference standard and lessen the burden to patients.
Getting from evidence to recommendations

• Recommendations are judgments about
  – Trade off between benefits and harms
  – Quality of evidence
  – Values and preferences
  – Resource use
## Domains that contribute to the strength of a recommendation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance between desirable and undesirable outcomes (estimated effects)</td>
<td>The larger the differences between desirable and undesirable consequences, the more likely a strong recommendation is warranted</td>
</tr>
<tr>
<td>Overall quality of evidence for outcomes</td>
<td>The higher the quality of evidence, the more likely a strong recommendation is warranted</td>
</tr>
<tr>
<td>Confidence in values and preferences and variability</td>
<td>The greater the variability (or uncertainty) in values and preferences, the more likely a weak recommendation is warranted</td>
</tr>
<tr>
<td>Resource use</td>
<td>The higher the cost and the more resources consumed, the less likely a strong recommendation is warranted</td>
</tr>
</tbody>
</table>
Strength of recommendations

• The strength of a recommendation is defined as the extent to which one can be confident that the desirable effects of an intervention outweigh its undesirable effects
  - Strong for
  - Strong against
  - Weak for
  - Weak against

Weak is also called conditional, discretionary, or qualified

Andrews JC, J Clin Epi 2013
What GRADE means by strong and weak recommendations for clinicians and patients

• Strong - all or almost all informed people would make the recommended choice for or against the recommended course of action (uniformity)

• Weak - most informed people would choose the recommended course of action, but a substantial number would not (variability)

• Also, consider prevalence, equity, cost, and improving quality of care
## Implications of strong and weak recommendations

<table>
<thead>
<tr>
<th></th>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For patients</strong></td>
<td>Most individuals in this situation would want the recommended course of action and only a small proportion would not</td>
<td>The majority of individuals in this situation would want the suggested course of action, but many would not</td>
</tr>
<tr>
<td><strong>For clinicians</strong></td>
<td>Most individuals should receive the recommended course of action</td>
<td>Different choices will be appropriate for different patients…consistent with values/prefences</td>
</tr>
<tr>
<td><strong>For policy makers</strong></td>
<td>Can be adopted as policy in most situations</td>
<td>Policy making will require debate and involvement of many stakeholders</td>
</tr>
</tbody>
</table>
Systematic review

Guideline development

Formulate question, then select outcomes. Rate importance.

- Critical
- Important
- Not important

Create evidence profile with GRADEpro.

Rate quality of evidence for each outcome.

- High
- Moderate
- Low
- Very low

Summary of findings & estimate of effect for each outcome.

- Large effect
- Dose response

Rate overall quality of evidence across outcomes based on lowest quality of critical outcomes.

We recommend using...
We suggest using...
We recommend against using...
We suggest against using...

By considering:
- Quality of evidence
- Balance benefits/harms
- Values and preferences
- Resource use (cost)

Revise if necessary by considering:

Panel
Challenges

Image, Copyright Suzanne Beaky
57% of TB recommendations were strong based on low/very low quality evidence
Four themes emerged

• High standards essential for credibility
• Mixed views on need for single quality assurance process (set by the Guidelines Review Committee)
• Uncertainties about applying GRADE
• Technical capacity to implement new standards is variable

“Since 2007, WHO guideline development methods have become more systematic and transparent. However, some departments are bypassing the procedures, and as yet neither the Guidelines Review Committee, nor the quality assurance standards they have set, are fully embedded within the organization.”
Canadian guideline development

730 guidelines 1994 to 1999; 630 from 2000 to 2005

• Guidelines produced more recently in Canada are less likely to be based on a review of the evidence
• Only half discuss levels of evidence underlying recommendations
• More frequent producers were more likely to have conducted a computerized literature search (91.6% compared to 76.6%) and graded the quality of evidence (51.6% compared to 38.1%)

Implementation Science

Research article

Twelve years of clinical practice guideline development, dissemination and evaluation in Canada (1994 to 2005)
Jennifer Kryworuchko¹, Dawn Stacey¹, Nan Bai² and Ian D Graham*¹,³

Open Access
Challenges applying GRADE

- GRADE requires considerable training and experience
- The identical body of evidence can be appraised differently by judges with different individual biases or values
- Insufficient for making informed decisions about scale-up
Challenges applying GRADE to diagnostic tests

Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. It made a difference whether assessors looked at the evidence from a patient-important outcome perspective or a test accuracy perspective. Inconsistency, imprecision, and publication bias were challenging to apply.

On the issue of representativeness of patient populations, assessors had to be conscious to not downgrade the evidence twice for criteria “risk of bias” and “indirectness”.

A clear distinction is needed between test accuracy and patient-important outcomes as the outcome included impacts judgments about evidence quality.
Further development in GRADE for diagnostics

• The “PICO” format needs to be adapted for diagnostic questions (e.g., define test-treatment pathway in both diagnostic test accuracy studies as well as in test accuracy reviews)
• More guidance is needed in applying the GRADE criteria for imprecision, inconsistency, and publication bias
• Default downgrading of test accuracy evidence on the basis of indirectness to patient outcomes is not always necessary (examples are needed)

Gopalakrishna et al. J Clin Epi 2014
Which New Diagnostics for Tuberculosis, and When?

Frank Cobelens,1 Susan van den Hof,1,2 Madhukar Pai,3 S. Bertel Squire,4 Andrew Ramsay,5 and Michael E. Kimerling6 on behalf of the Evidence for Scale-up Group

Figure 3. Proposal for a revised pathway focused on the postaccuracy phase of tuberculosis diagnostics, showing the proposed value chain for new diagnostic tests for tuberculosis. The grey arrows in the middle represent the stages in the evaluation pathway, and the colored boxes represent policy decisions at the global level (red) and the country level (blue). Countries would adopt implementation at different points and should provide feedback about their experiences (* in the blue box). In the stages before scale-up and during and after scale-up, evaluation data would be collected on diagnostic algorithms incorporating the new test.
Standards for guidelines

1. Establishing transparency
2. Management of conflict of interest
3. Guideline development group composition
4. Clinical practice guideline-systematic review intersection
5. Establishing evidence foundations for and rating strength of recommendations
6. Articulation of recommendations
7. External review
8. Updating

Acknowledgements

• Holger Schünemann

• Nancy Santesso
References


8. GRADE Working Group gradeworkinggroup.org/

9. GRADE learning modules cebgrade.mcmaster.ca

10. GRADE handbook http://www.guidelinedevelopment.org/handbook/#h.j8tx801skdu3