

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



Haynes. BMJ 2002;324:1350

Which hierarchy? (1)

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

Evidence	Recommendation	Organization
• B	Class I	➤ AHA
• A	1	➤ ACCP
• IV	C	➤ SIGN

Centers for Disease Control and Prevention (CDC)

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

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



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 BMJ VOLUME 327 20–27 DECEMBER 2003

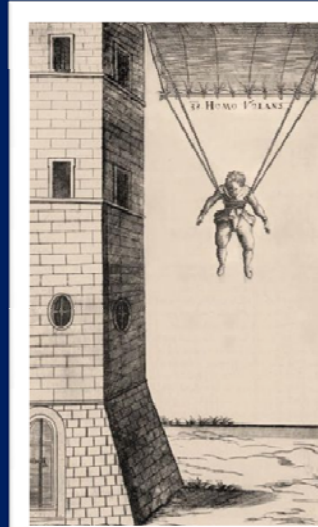


Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials

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

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

CMAJ 2003, BMJ 2004, BMC 2004, BMC 2005, AJRCCM 2006, Chest 2006, BMJ 2008

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

15

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



Belief ≠ confidence

Figure 1. Belief and confidence: a two-dimensional weather report. (Reprinted by permission from the Wall Street Journal).

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

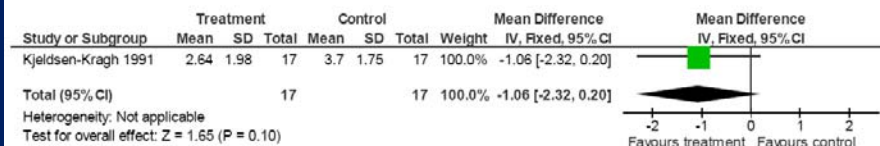


- There always is evidence
“When there is a question there is evidence”
- Better research \Rightarrow greater confidence in the evidence and decisions

Controlled trial of fasting and one-year vegetarian diet in rheumatoid arthritis. The Lancet; 338: 899-902

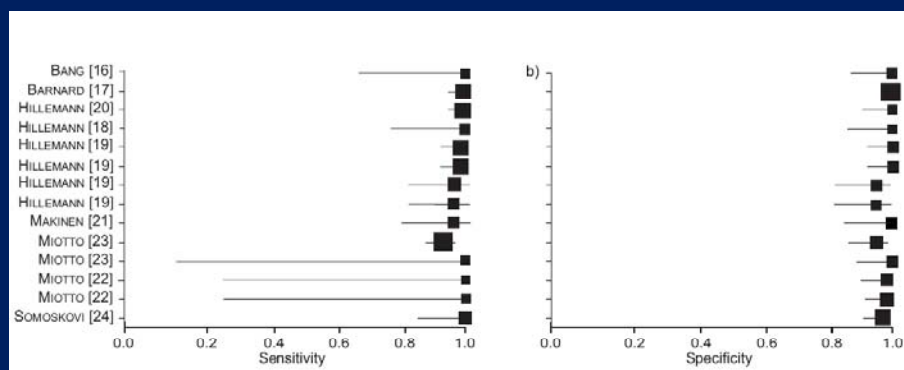
1 7-10 days fasting followed by vegetarian diet vs ordinary diet

1.1 Pain (0-10) 3 weeks follow up



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

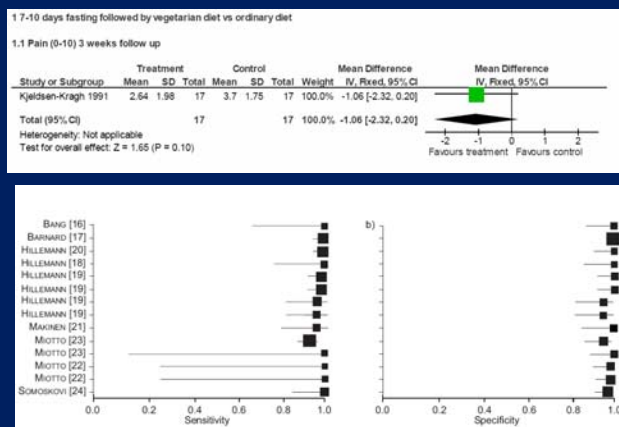
GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis, 2008



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?



RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies

The GRADE system can be used to grade the quality of evidence and strength of recommendations for diagnostic tests or strategies. This article explains how patient-important outcomes are taken into account in this process

In this fourth article of the five part series, we describe how guideline developers are using GRADE to rate the quality of evidence and move from evidence to a recommendation for diagnostic tests and strategies. Although recommendations on diagnosis share the fundamental logic of recommendations for other interventions, they present unique challenges. We will describe why guideline panels should be cautious when they use evidence of the accuracy of tests ("test accuracy") as the basis for recommendations and why evidence of test accuracy often provides low quality evidence for making recommendations.

Hölger J Schünemann professor,
Department of Epidemiology,
Italian National Cancer Institute
Regina Elena, 00144 Rome, Italy
and CLARITY Research Group,
Department of Clinical Epidemiology
and Biostatistics, McMaster
University, Hamilton, Ontario,
Canada L8N 3Z5
Andrew D Osman researcher,
Norwegian Knowledge Centre for
the Health Services, PO Box 7004,
0130 Oslo, Norway
Jan Brozek research fellow,
Department of Epidemiology, Harvard

outcome. For patients who present with apparently operable lung cancer, the presumption is that additional tests will spare patients the morbidity and early mortality associated with futile thoracotomy. The example of computed tomography for coronary artery disease in the box illustrates another common rationale for a new test: replacement of another test (coronary computed tomography instead of conventional angiography) to avoid complications associated with a more invasive and expensive alternative.⁶

The best way to assess any diagnostic strategy—but in particular new strategies with putative superior

Schünemann. BMJ 2008;336:1106-1110; www.gradeworkinggroup.org/

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?

	Representative spectrum?	Acceptable reference standard?	Acceptable delay between tests?	Partial verification avoided?	Differential verification avoided?	Incorporation avoided?	Reference standard results blinded?	Index test results blinded?	Relevant clinical information?	Uninterpretable results reported?	Withdrawals explained?	Conducted without industry involvement?
Al-Orainey 1992a	+	+	+	+	+	+	?	+	+	?	?	+
Al-Orainey 1992b	+	+	+	+	+	+	?	+	+	?	?	+
Alavi-Naini 2009a	+	+	+	+	+	+	+	+	+	?	+	+
Alavi-Naini 2009b	+	+	+	+	+	+	+	+	+	?	+	+
Araj 1993a	+	+	+	+	+	+	?	?	+	?	?	+

Methodological quality summary: review authors' judgments about each methodological quality item for each included study, created in RevMan <http://ims.cochrane.org/revman>

Who believes the risk of bias is of concern?

Yes

No

Don't know or undecided

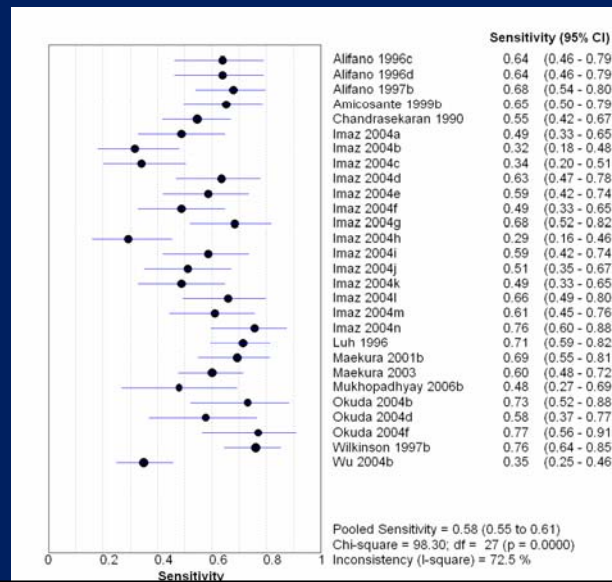
Would you downgrade for risk of bias?

- ☐ 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



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 3. GRADE Summary of Findings. Should commercial serological tests be used as a replacement test for conventional tests such as smear microscopy in patients of any age suspected of having pulmonary tuberculosis?

Outcome	No. studies (Participants)	Study Design	Limitations	Indirectness	Inconsistency	Imprecision	Publication bias	Final Quality	Effect per 1000	Importance
True Positives	67 (8318)	Cross-sectional and case-control	Very Serious ^{A1} (-2)	No Serious Indirectness ^{A2}	Very Serious ^{A3} (-2)	Serious ^{A4}	Likely ^{A5}	Very Low ⊕○○○	Prevalence 10%; 64 Prevalence 30%; 192	Critical
True Negatives	67 (8318)	Cross-sectional and case-control	Very Serious ^{A1} (-2)	No Serious Indirectness ^{A2}	Very Serious ^{A3} (-2)	Serious ^{A4}	Likely ^{A5}	Very Low ⊕○○○	Prevalence 10%; 819 Prevalence 30%; 637	Critical
False Positives	67 (8318)	Cross-sectional and case-control	Very Serious ^{A1} (-2)	No Serious Indirectness ^{A2}	Very Serious ^{A3} (-2)	Serious ^{A4}	Likely ^{A5}	Very Low ⊕○○○	Prevalence 10%; 81 Prevalence 30%; 63	Critical
False Negatives	67 (8318)	Cross-sectional and case-control	Very Serious ^{A1} (-2)	No Serious Indirectness ^{A2}	Very Serious ^{A3} (-2)	Serious ^{A4}	Likely ^{A5}	Very Low ⊕○○○	Prevalence 10%; 36 Prevalence 30%; 108	Critical

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

^{A1}Majority of studies lacked a representative patient spectrum and were not blinded.

^{A2}Although diagnostic accuracy is considered a surrogate for patient-important outcomes, we did not downgrade.

^{A3}There was considerable heterogeneity in study results.

^{A4}We did not pool accuracy estimates. The 95% CIs were wide for many individual studies. We did not downgrade as there were a large number of studies and we already took off 2 points for inconsistency.

Footnotes

Steingart et al, submitted manuscript

False Positives	67 (8318)	Cross-sectional and case-control	Very Serious ^{A1} (-2)	No Serious Indirectness ^{A2}	Very Serious ^{A3} (-2)
False Negatives	67 (8318)	Cross-sectional and case-control	Very Serious ^{A1} (-2)	No Serious Indirectness ^{A2}	Very Serious ^{A3} (-2)

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

^{A1}Majority of studies lacked a representative patient spectrum and were not blinded.

^{A2}Although diagnostic accuracy is considered a surrogate for patient-important outcomes, we did not

^{A3}There was considerable heterogeneity in study results.

^{A4}We did not pool accuracy estimates. The 95% CIs were wide for many individual studies. We did not

2 points for inconsistency.

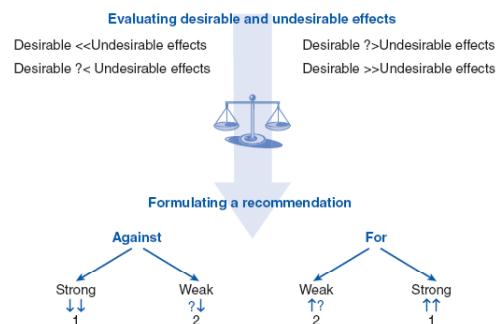
*Moving From Evidence to
Recommendations*

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

Strength of 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

Guyatt. BMJ, 10 May 2008, Volume 336:1049-51


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

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

Recommendation: In patients with HIV and drug resistant TB requiring second line drugs, the expert panel recommends/suggests to (not) administer ART (? recommendation, ? quality evidence).			
Population: HIV positive individuals with drug resistant TB requiring second line drugs			
Intervention: ART use during TB treatment versus ART non-use			
Factor	Decision	Explanation	
High or moderate quality evidence (Is there high quality evidence?)	<input type="checkbox"/> Yes <input type="checkbox"/> No	 There is limited evidence from published studies to evaluate ART use in HIV-TB coinfecting patients receiving second line drugs for XDR-TB and MDR-TB.	
Certainty about the balance of benefits versus harms and burdens (Is there certainty?)	<input type="checkbox"/> Yes <input type="checkbox"/> No	Although there is some uncertainty about cure, there is a significant decrease in hazards ratio for death even after controlling for initial CD4 count	
Certainty or similarity in values (Is there certainty?)	<input type="checkbox"/> Yes <input type="checkbox"/> No	Cure and survival appear to be more likely in drug resistant TB requiring second line drugs if ART is used during TB treatment. HR of 3.17 (1.46, 6.9) for cure and HR of 0.41 (0.26, 0.63) for death in ART vs. non ART group	
Resource implications (are the resources consumed worth the expected benefit?)	<input type="checkbox"/> Yes <input type="checkbox"/> No	More resources required for concomitant ART use	Little uncertainty regarding the outcomes of cure and survival. Significant uncertainty regarding effects of ART on other outcomes, including adverse events, default, time to smear, culture conversion, timing of ART initiation Need for more skilled providers trained in HIV and drug resistant TB care and drug-drug interactions Need for increased integration of HIV and TB care
Overall strength of recommendation	Strong or conditional		

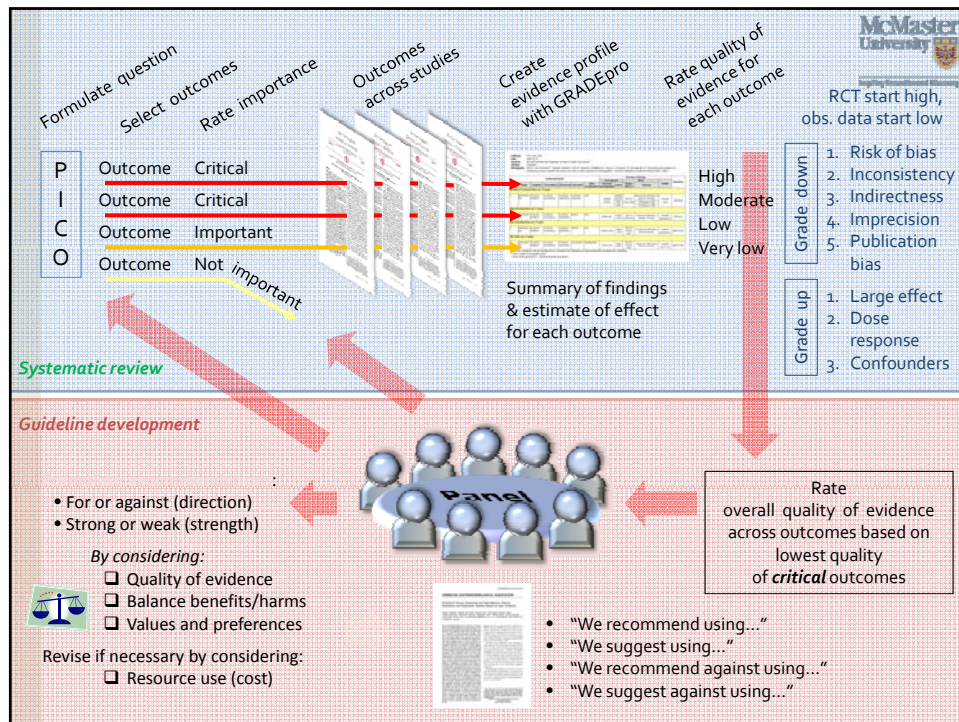
Matt Arentz et al, unpublished

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

WHO TB diagnostics policy formulation process





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*