Grading Diagnostic Evidence-Based Statements and Recommendations

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

GRADE

www.gradeworkinggroup.org

Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Williams J, et al. GRADEing the quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* submitted.

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[r]
*Holger J Schünemann, associate professorg[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[v]
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**

Recommendations about Diagnosis

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.

Grading Recommendations:

Strong or Weak (For or Against)

Strong recommendations

- strong methods
- accurate test impacts outcome
- few downsides of testing strategy
- indicating a judgment that a majority of well informed people will make the same choice
  - (high confidence, low uncertainty)
- expect non-variant clinician and patient behavior - most patients should receive the intervention
  - diminished role for clinical expertise - focus on implementation & barriers
  - focused role of patient values and preferences - emphasis on compliance and barriers
  - could be used as a performance / quality indicator
- decision aids not likely to be needed
- medical practice is expected to not to vary much

Weak recommendations

- weak methods
- imprecise estimate / small effect
- substantial downsides
- indicating a judgment that a majority of well informed people will make the same choice, but a substantial minority will not (significant uncertainty)
- expect variability in clinician and patient actions
  - clinical expertise important - focus on decision-making and implementation
  - patient values and preferences important - focus on determining values and preferences relative to decision
- decision aids likely to be useful
  - offering the intervention and helping patients make a decision could be used a quality criterion
- medical practice is expected to vary to some degree

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)

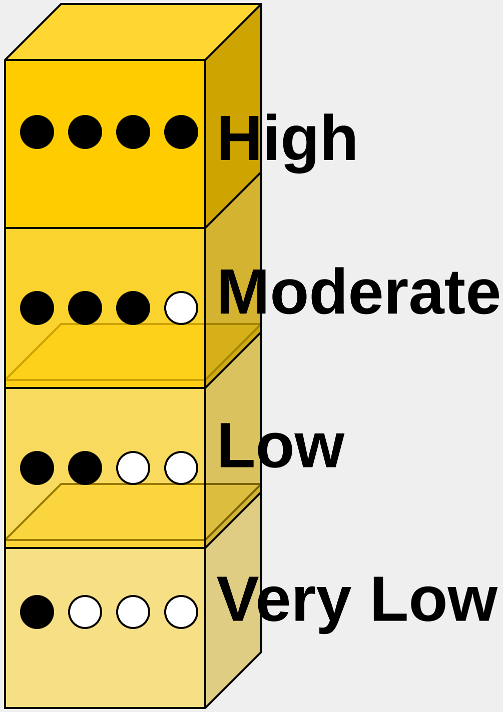
Grading evidence process

Describe question / problem (PICO)
Establish critical / important outcomes to decision and relative importance of outcomes
Systematic review
Evidence profile / quality for each outcome
Overall quality of evidence
Benefit/risk/harm/cost balance
Overall strength of recommendation - GRADE
Implementation and evaluation

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

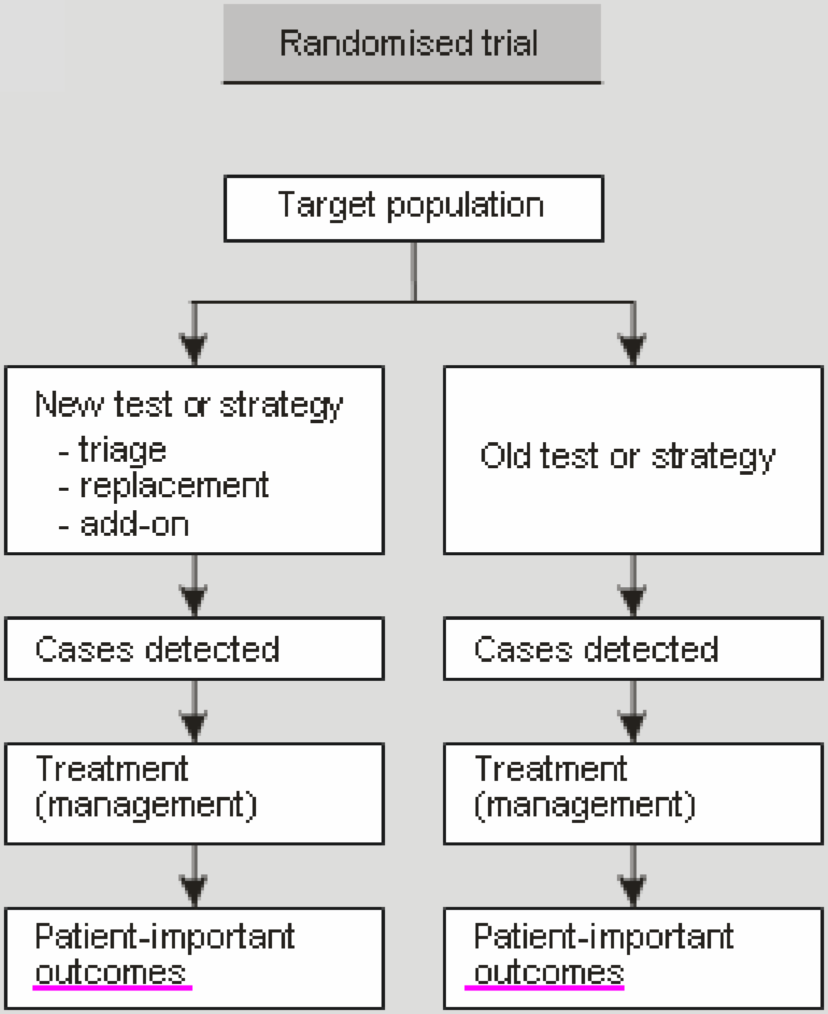


Starting Line

Blinded testing of previously developed diagnostic criteria in series of consecutive patients (with universally applied reference "gold" standard) or a systematic review of these studies
Development of diagnostic criteria on basis of consecutive patients (with universally applied reference "gold" standard) or a systematic review of these studies
Study of nonconsecutive patients (no consistently applied reference "gold" standard) or a systematic review of these studies
Study with a poor reference standard / Case-control study / Expert opinion / Case Report

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

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

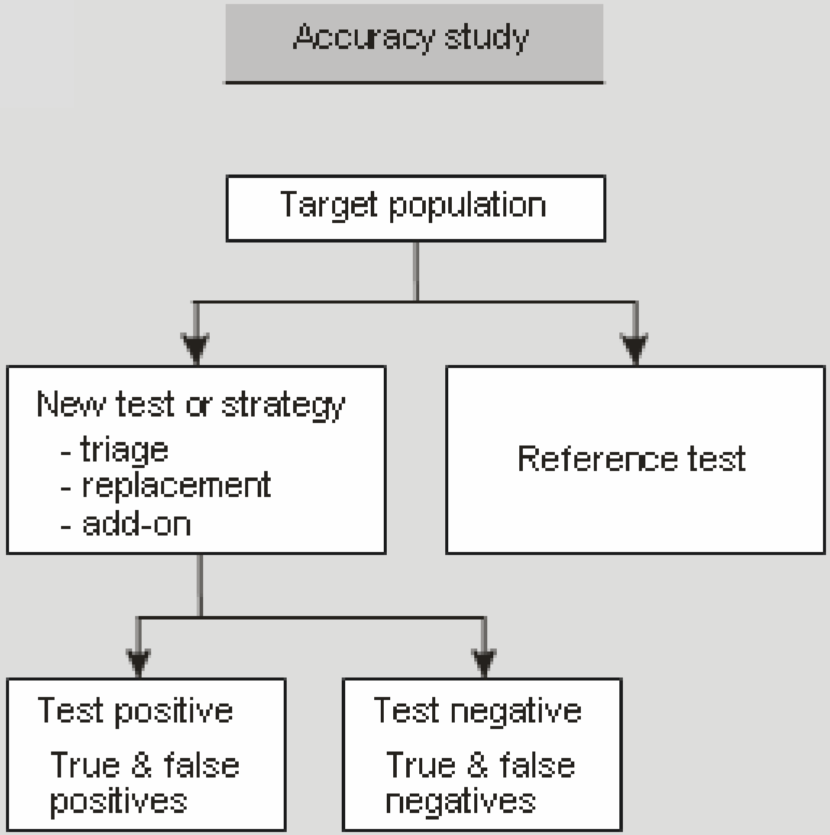


- Example: RCT that explored a diagnostic strategy guided by the use of B-type natriuretic peptide (BNP)
- Designed to provide a more accurate diagnosis of heart failure - in patients presenting to the emergency department with acute dyspnea.
- The group randomized to receive BNP spent a shorter time 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.

Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med* 2004;350(7):647-54.

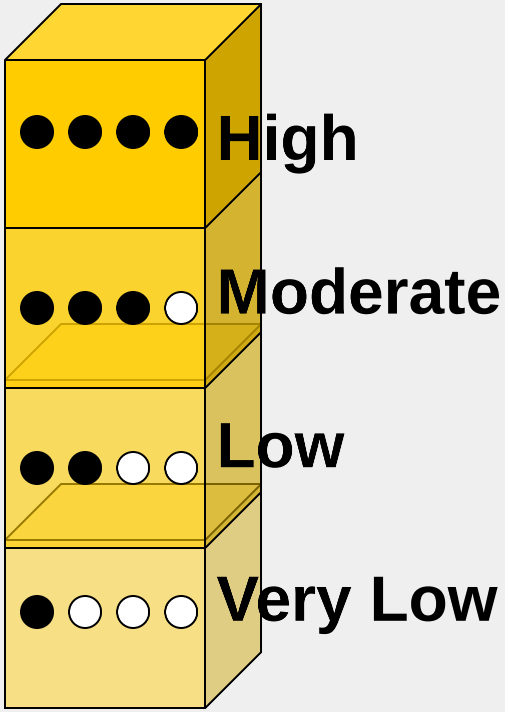
Types of studies to evaluate a test or diagnostic strategy: accuracy-based

- 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.



Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *Bmj* 2001;323(7305):157-62.
Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med* 2002;40(3):280-6.
Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? *Can Assoc Radiol J* 2002;53(3):144-8.

Judgments about Evidence Quality



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

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

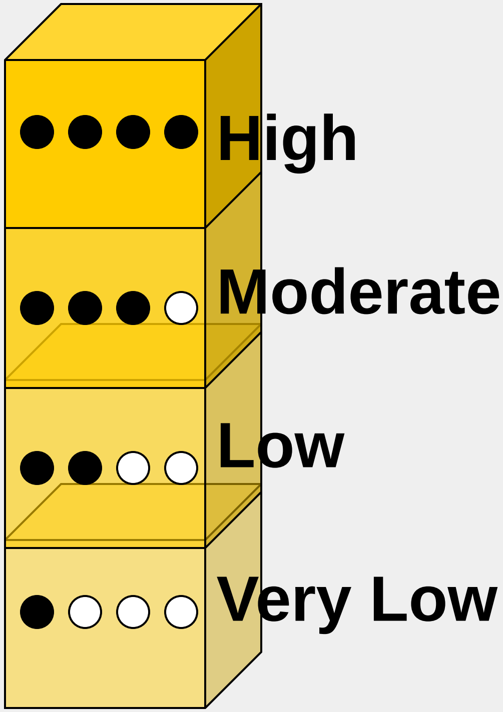
- selection bias or reporting bias: strong evidence lower by 1 level

- 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

Coronary CT scanning versus invasive angiography: arriving at a bottom line for study quality

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]

All outcomes

No of studies	**Design**	Limitations	Directness	Inconsistency	Imprecise data	Reporting Bias	Quality
21 studies (1,570 patients)	accuracy studies[1]	no	Strong relation to patient important outcomes	yes[2] (-1)	no	no[3]	⊕⊕⊕○ Moderate

[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%)

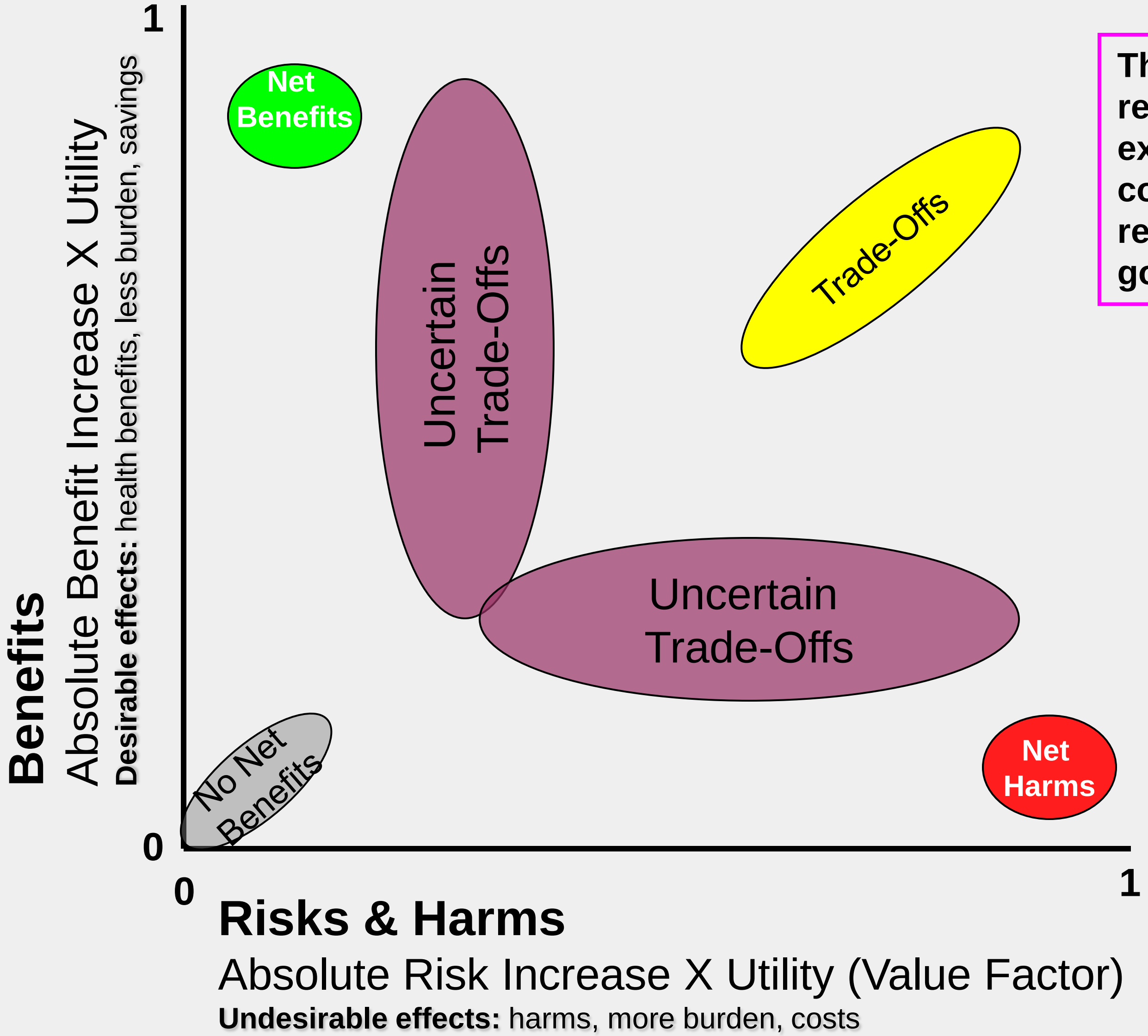
[2] there was significant heterogeneity of results (not investigated further by the authors of the review)

[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


1
0
0
1
Net Benefits
Uncertain Trade-Offs
Trade-Offs
Uncertain Trade-Offs
No Net Benefits
Net Harms
Benefits
Absolute Benefit Increase X Utility
Desirable effects: health benefits, less burden, savings
Risks & Harms
Absolute Risk Increase X Utility (Value Factor)
Undesirable effects: harms, more burden, costs

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.

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.

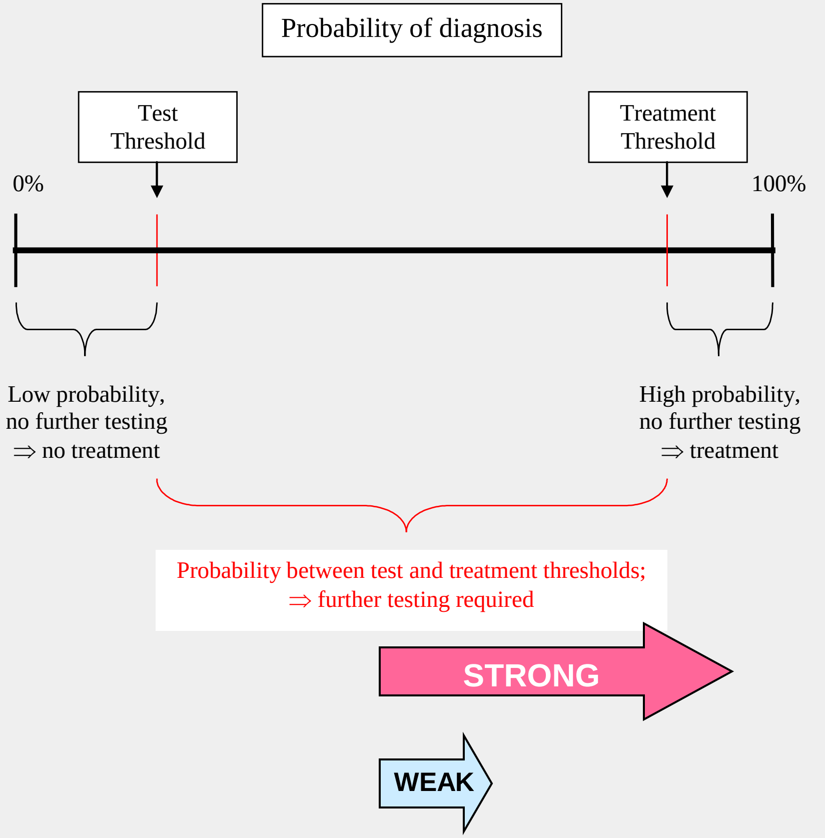
Net harms: The test intervention does more harm than good.

Decision Thresholds and Diagnostic Tests: Alternative Conceptualization

- 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.



Probability of diagnosis
Test Threshold
Treatment Threshold
0%
100%
Low probability, no further testing ⇒ no treatment
High probability, no further testing ⇒ treatment
Probability between test and treatment thresholds; ⇒ further testing required
STRONG
WEAK

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.

Citations

- Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328(7454):1490.
- Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. Am J Respir Crit Care Med 2006;174(5):605-14.
- Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Williams J, et al. GRADEing the quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ submitted.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003;138(1):40-4. http://www.stard-statement.org/website%20stard/
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138(1):W1-12. http://www.stard-statement.org/website%20stard/
- Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332(7549):1089-92.
- Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061-6.
- Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469-76.
- Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25. http://www.biomedcentral.com/1471-2288/3/25
- Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9. http://www.biomedcentral.com/1471-2288/6/9