Grading evidence from test accuracy studies: what makes it challenging compared with the grading of effectiveness studies?

Ewelina Rogozińska,[1,2] Khalid Khan[1,2]

10.1136/ebmed-2017-110717

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/ebmed-2017-110717)

[1]Women's Health Research Unit, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK
[2]Department of Multidisciplinary Evidence Synthesis Hub (mEsh), Barts and the London School of Medicine, Queen Mary University London, London, UK

Correspondence to:
Ewelina Rogozińska, Women's Health Research Unit, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK;
e.a.rogozinska@qmul.ac.uk

Abstract

Guideline panels need to process a sizeable amount of information to issue a decision on whether to recommend a health technology or not. Grading of Recommendations Assessment, Development, and Evaluation (GRADE) is being frequently applied in guideline development to facilitate this task, typically for the synthesis of effectiveness research. Questions regarding the accuracy of medical tests are ubiquitous, and they temporally precede questions about therapy. However, literature summarising the experience of applying GRADE approach to accuracy evaluations is not as rich as one for effectiveness evidence. Type of study design (cross-sectional), two-dimensional nature of the performance measures (sensitivity and specificity), propensity towards a higher level of between-study heterogeneity, poor reporting of quality features and uncertainty about how best to assess for publication bias among other features make this task challenging. This article presents solutions adopted to addresses above challenges for judicious estimation of the strength of test accuracy evidence used to inform evidence syntheses for guideline development.

Introduction

Grading of Recommendations Assessment, Development and Evaluation (GRADE) is being increasingly used to synthesise evidence for practice and policy development.[1 2] The GRADE domains, that is, type of evidence and its consistency, directness, precision and risk of bias, etc,[3 4] are frequently and readily applied to therapeutic effectiveness research.[5 6] However, clinical practice requires direction about the accuracy of tests to make a diagnosis before contemplating over decisions about treatment.[7] For assessment of evidence concerning the former, guidance on the use of GRADE principles still requires more attention.[4 8] The aim of this paper is to raise awareness of grading the strength of test accuracy evidence, associated with its challenges, and contrasting them with the issues relevant for the evaluation of effectiveness research. We use grading of the quality of test accuracy evidence employed in a WHO guidelines on antenatal care for a positive pregnancy experience[9] as exemplary.

The basics: accuracy versus effectiveness research

Typically in test accuracy research, the question format is as follows: clearly defined participants, an object of the evaluation (an index test) and a comparator (a reference standard test to verify the presence or absence of outcome or condition of interest) (table 1). The 2×2 contingency table created this way can be used to calculate test accuracy measures such as sensitivity and specificity.[10] Accuracy research informs us about how well tests can detect given a condition. In conjunction with effectiveness research it can be used to inform an antenatal management algorithm to rationalise the use of tests and treatments. If the effectiveness of interventions is unclear or unknown, assessment of test accuracy has limited utility. Equally, if accurate tests do not exist, it is difficult to know whom to treat. Whereas the definitive study design for effectiveness research is a controlled trial with randomisation,[11] study designs for evaluation of test accuracy do not require this approach. The most valid accuracy results are obtained from cross-sectional studies that concurrently apply index and reference tests and avoid features that can introduce bias.[12]

Further in the text, to illustrate the application of GRADE approach to accuracy research, we used an example (table 2) derived from the assessment prepared to inform the WHO antenatal guideline.[9] The guideline was prepared in line with the WHO internal standards and guided by standard operating procedures both authors took part in developing (details available on request). Undetected asymptomatic bacteriuria, if left untreated in pregnancy, might lead to serious complications,[13] and the quality of accuracy evidence for urine dipstick (nitrites marker only) in detecting the infection was one of the evaluations prepared for the guideline (figure 1). Details of the full evaluation of the accuracy of on-site tests to detect asymptomatic bacteriuria are available elsewhere.[14] Robustness of all GRADE features (table 1) was considered for their potential to weaken the overall strength of evidence through downgrading of individual aspects.

Risk of bias

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool is used for the assessment of the risk of bias in the accuracy evidence. The tool comprises domains that can be assessed as a low, unclear or high risk of bias for participant selection, implementation of the index test, the reference standard and study flow and timing.[15] The approach is based on the same concept as the tool used to assess the effectiveness research[16] with domains relevant to study design used in accuracy research.

The accuracy evidence for urine dipstick was downgraded from 'not serious' to 'serious' (table 2), as more than a half of the pooled studies was classified as the moderate or high risk of bias (see online supplementary appendix 1). Before grading, the studies were classified as low, moderate or high of a risk of bias based on the

Table 1 Differences between grading of strength of accuracy and effectiveness evidence

Items	Effectiveness research	Accuracy research	Issues to consider
The basics			
Question	Participants, Intervention, Comparator and Outcome(s)	Population, Index test and Reference standard	PICO structure does not readily apply in diagnostic research
Study design	Randomised controlled trial (RCT) Non-randomised controlled studies	Cross-sectional Other designs for index test and reference standard comparison	RCT not required for accuracy evaluations
Measure of performance	Effect estimate (OR or risk ratio)	Accuracy estimates (sensitivity and specificity)	Accuracy estimates are usually paired; global single accuracy measures are not very intuitive
GRADE feature			
Risk of bias	Tools for the risk of bias typically assess: random sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting	Methodological quality part QUADAS-2 instrument includes: participant selection, implementation of the index test, reference standard, flow and timing	Various tools exist; there is no consensus on what is best for accuracy research
Indirectness	Based on PICO question (see above)	QUADAS-2 instrument applicability part can be deployed	Accuracy research requires looking at participants, reference standard, flow and timing
Inconsistency	I^2 or χ^2 tests for heterogeneity	Visual assessment of overlap of CIs between studies	Assessed separately for sensitivity and specificity in accuracy research
Imprecision	95% CIs around an effect measure	95% CIs around multiple accuracy measures	Assessed separately for sensitivity and specificity in accuracy research
Publication bias	Test for funnel plot asymmetry	Test for funnel plot asymmetry	Test for funnel asymmetry requires particular caution in accuracy research

PICO, Population Intervention Comparator Outcome; GRADE, Grading of Recommendations Assessment, Development and Evaluation; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies.

Table 2 GRADE assessment of evidence quality of urine dipstick (nitrites) accuracy (index test) to detect asymptomatic bacteriuria (reference standard: urine culture) in pregnancy[8]

Outcome (presence or absence of bacteriuria)	No of studies (No of patients)	Accuracy measures	Features that may affect evidence quality*					Quality of evidence‡
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias†	
True positives (patients with asymptomatic bacteriuria) False negatives (patients incorrectly classified as not having asymptomatic bacteriuria)	21 cross-sectional studies (699 patients)	Sensitivity 0.56 (95% CI 0.42 to 0.69)	Serious§	Serious¶	Serious**	Serious††	–	⊕◯◯◯ Very low
True negatives (patients without asymptomatic bacteriuria) False positives (patients incorrectly classified as having asymptomatic bacteriuria)	21 cross-sectional studies (8560 patients)	Specificity 0.99 (95% CI 0.98 to 0.99)	Serious§	Serious¶	Not serious	Not serious	–	⊕⊕◯◯ Low

*As specified in table 1.

†Domain was not assessed (see text for justification).

‡See figure 1 for graphic display of quality of evidence.

§57% of included studies of low and moderate quality[14] (in the cited reference the detail quality assessment appears in appendix 5).

¶54% of included studies were assessed as 'high' or 'unclear' concern over applicability.

**Visible variability between studies on the forest plot.

††Noticeable imprecision, wide CIs.

respective scoring of the domains (see online supplementary appendix 2).

Indirectness

The QUADAS-2 tool comprises two parts: the first focusing on the methodological quality of the study design (discussed above), and the second addressing the applicability of the study to the research question. The applicability part constitutes three domains allowing us to assess the indirectness of evidence with regards to population, reference standard and study flow and timing (see online supplementary appendix 1). For effectiveness research, the respective aspect is assessed basing on how well the populations, interventions,

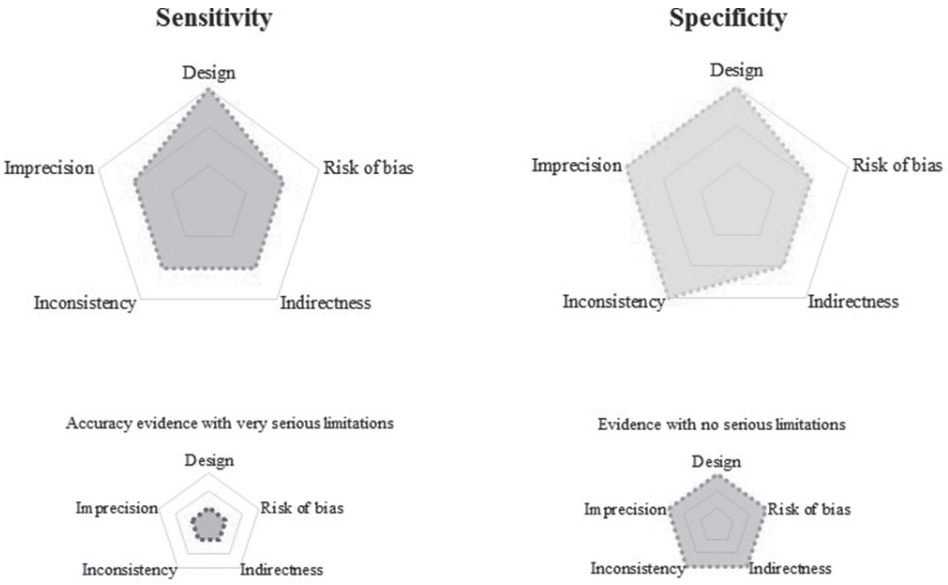

Figure 1 Graphic display of evidence quality of urine dipstick (nitrites marker only) accuracy to detect asymptomatic bacteriuria in pregnancy (top graphs). The graphs represent the quality of features shown on the shape corners. For each of the corners, the distance from the centre represents the level of evidence strength with the lowest close to the shape's centre (bottom left example) and highest at its maximum (bottom right graph).

comparators and reported outcomes match the research question. The QUADAS-2 tool, with its applicability part, allows assessing the indirectness of accuracy studies in a more structured and transparent way than it is being done for effectiveness research. We set a grading rule for applicability of synthesised evidence (see online supplementary appendix 1) that leads us to the downgrading of the evidence strength in our example as around 50% of the studies used in the synthesis was assessed as 'high' or 'unclear' concern over their applicability (table 2).

Inconsistency

Between-study heterogeneity is anticipated more often for accuracy than effectiveness research. Furthermore, the potential inconsistency can occur not for one but two performance measures (table 1). Grading of the accuracy evidence is two dimensional with its strength assessed separately for sensitivity and specificity. The test used to evaluate between-study heterogeneity used in the effectiveness research does not work well for accuracy in this case. We chose to assess the inconsistency between the accuracy measures through visual inspection of the overlap of CI around the performance measures between pooled studies. The domain was graded depending on the degree of lack of overlap between CIs (see online supplementary appendix 1). The evidence for the sensitivity of urine dipstick (nitrites) was downgraded to seriously decreasing the quality of evidence due to visible variability in the performance estimates between the studies[14] (table 2).

Imprecision

The wider the CI of pooled estimates, the poorer the precision and the weaker the strength of evidence. When grading the imprecision of performance measures, the same rule applies to both types of research with a similar challenge when the occurrence of the condition (event) is rare. If the prevalence of the condition is low, CIs around the pooled performance measure are wide. Due to the dual nature of the accuracy performance measure, we observe that the CI for pooled sensitivity tends to be wider than for the pooled specificity. The consequence of this is a differential assessment of the evidence strength for test sensitivity and specificity as in our example (table 2).

Publication bias

Funnel plot asymmetry tests are used to examine the impact of the effects from small studies and are being treated as an indicator of potential risk of publication bias.[17]

A statistical test taking into account effective sample size and associated regression statistical test of asymmetry for detection of sample size-related bias are currently recommended when pooling accuracy studies.[18] In comparison to the statistical tests that use SEs of ORs, commonly used in the effectiveness research, that are likely to be misleading if applied to a meta-analysis of the accuracy measures. However, the impact of small-study effects is not as clear in accuracy research, and the power of the currently available test is modest[19] leading us to a decision to leave out this domain (table 2).

Conclusion

Accuracy research as an important element of any clinical management algorithm requires a thorough and unequivocal assessment of its quality for evidence syntheses. While assessment of domains such risk of bias, indirectness or impression of accuracy measures in the evidence synthesis should not pose any greater challenges than in the case effectiveness research, more insight is needed into the impact of the heterogeneity and the publication bias on the synthesis of accuracy evidence to facilitate this task.

Without a doubt, members of the GRADE Working Group are aware of the above-mentioned issues and in due course will surely see more guidance on the application of GRADE to accuracy evidence with our work contributing to its use. Hopefully, the future guidance will also cover application of GRADE to evidence derived from a single study and use of likelihood ratio as a parameter describing test performance generally better understood by the clinicians.[20]

Acknowledgements The authors would like to acknowledge the assistance of the following advisors from the WHO Department of Reproductive Health and Research: A Metin Gülmezoglu, Özge Tunçalp and Professor Javier Zamora from Clinical Biostatistics Unit, Hospital Ramon y Cajal (IRYCIS) and CIBER Epidemiology and Public Health, Madrid.

Contributors KK and ER prepared the diagnostic GRADE assessment for the evidence on the accuracy of the onsite test to detect asymptomatic bacteriuria in pregnancy in GRADEpro GDT (web). ER wrote the initial draft of the manuscript and all subsequent drafts after critical review by KK. KK is guarantor for the manuscript.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

References

1. **Developing NICE Guidelines: The Manual.** *NICE process and methods guides*. London: National Institute for Health and Care Excellence, 2015.
2. **World Health Organization.** Evidence retrieval and synthesis. *WHO handbook for guideline development*. 2nd ed, 2014:93–108.
3. **Guyatt GH**, Oxman AD, Kunz R, *et al*. Going from evidence to recommendations. *BMJ* 2008;**336**:1049–51.
4. **Schünemann HJ**, Schünemann AH, Oxman AD, *et al*. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;**336**:1106–10.
5. **Khan KS BE**, Roos C, Kowalska M, *et al*. For the EBM-CONNECT Collaboration. Making GRADE accessible: a proposal for graphic display of evidence quality assessments. *Evidence-Based Medicine* 2011;**16**.
6. **Murad MH**, Almasri J, Alsawas M, *et al*. Grading the quality of evidence in complex interventions: a guide for evidence-based practitioners. *Evid Based Med* 2017;**22**:20–2.
7. **Alonso-Coello P**, Schünemann HJ, Moberg J, *et al*. GRADE evidence to decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: introduction. *BMJ* **2016;353**:i2016.
8. **Hsu J**, Brożek JL, Terracciano L, *et al*. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. *Implement Sci* 2011;**6**:62.
9. **World Health Organization**. *WHO recommendations on antenatal care for a positive pregnancy experience*. Geneva: WHO Library, 2016.
10. **Leeflang MM**, Deeks JJ, Gatsonis C, *et al*. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;**149**:889–97.
11. **Khan KS**, Kunz R, Kleijnen J, *et al*. Five steps to conducting a systematic review. *J R Soc Med* 2003;**96**:118–21.
12. **Rutjes AW**, Reitsma JB, Di Nisio M, *et al*. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;**174**:469–76.
13. **Honest H**, Forbes CA, Durée KH, *et al*. Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health Technol Assess* 2009;**13**:1–627.
14. **Rogozińska E**, Formina S, Zamora J, *et al*. Accuracy of onsite tests to detect asymptomatic bacteriuria in pregnancy: a systematic review and meta-analysis. *Obstet Gynecol* 2016;**128**:495–503.
15. **Whiting PF**, Rutjes AW, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–36.
16. **Higgins JPT GSe**. Cochrane handbook for systematic reviews of interventions: the cochrane collaboration. 2011 cochrane-handbook.org.
17. **Sterne JA**, Sutton AJ, Ioannidis JP, *et al*. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002.
18. **Deeks JJ**, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;**58**:882–93.
19. **Macaskill PGC**, Deeks J, Harbord R, *et al*. Chapter 10 analysing and presenting results. *Cochrane handbook for systematic reviews of diagnostic test accuracy*, 2010.
20. **Whiting PF**, Davenport C, Jameson C, *et al*. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open* 2015;**5**:e008155.