Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable.

Gopalakrishna, G; Mustafa, Reem; Davenport, Clare; Scholten, Rob JPM; Hyde, CJ; Brozek, J; Schunemann, H; Bossuyt, PMM; Leeflang, MMG; Langendam, MW

DOI:
10.1016/j.jclinepi.2014.01.006

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Gopalakrishna, G, Mustafa, R, Davenport, C, Scholten, RJPM, Hyde, CJ, Brozek, J, Schunemann, H, Bossuyt, PMM, Leeflang, MMG & Langendam, MW 2014, 'Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable.', *Journal of Clinical Epidemiology*, vol. 67, no. 7, pp. 760-768. https://doi.org/10.1016/j.jclinepi.2014.01.006

Journal of Clinical Epidemiology 67 (2014) 760−768

Journal of Clinical Epidemiology

Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable☆

Gowri Gopalakrishna[a,*], Reem A. Mustafa[b,c,d], Clare Davenport[e], Rob J.P.M. Scholten[f], Christopher Hyde[g], Jan Brozek[b,c], Holger J. Schünemann[b,c], Patrick M.M. Bossuyt[a], Mariska M.G. Leeflang[a], Miranda W. Langendam[a]

[a]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands*

[b]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*

[c]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*

[d]*School of Medicine, University of Missouri-Kansas City 5100 Rockhill Road, Kansas City, MO 64110, USA*

[e]*Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Edgbaston, Birmingham B15 2TT UK*

[f]*Dutch Cochrane Center, University Medical Center, Postbus 85500 3508 GA Utrecht, The Netherlands*

[g]*Institute of Health Research, University of Exeter Medical School, The Veysey Building Salmon Pool Lane, Exeter EX2 4SG UK*

Accepted 14 January 2014; Published online 13 April 2014

Abstract

Objectives: The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group developed an approach to assess the quality of evidence of diagnostic tests. Its use in Cochrane diagnostic test accuracy reviews is new. We applied this approach to three Cochrane reviews with the aim of better understanding the application of the GRADE criteria to such reviews.

Study Design and Setting: We selected reviews to achieve clinical and methodological diversities. At least three assessors independently assessed each review according to the GRADE criteria of risk of bias, indirectness, imprecision, inconsistency, and publication bias. Two teleconferences were held to share experiences.

Results: For the interpretation of the GRADE criteria, it made a difference whether assessors looked at the evidence from a patient-important outcome perspective or from a test accuracy standpoint. GRADE criteria such as inconsistency, imprecision, and publication bias were challenging to apply as was the assessment of comparative test accuracy reviews.

Conclusion: The perspective from which evidence is graded can influence judgments about quality. Guidance on application of GRADE to comparative test reviews and on the GRADE criteria of inconsistency, imprecision, and publication bias will facilitate the operationalization of GRADE for diagnostics.

Keywords: GRADE; Cochrane diagnostic test accuracy systematic reviews; Diagnostic test accuracy; Diagnostics; Systematic reviews; Medical tests

1. Introduction

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group over the last 13 years developed a rigorous methodology for assessing the quality of the evidence and grading the strength of recommendations in health care [1−5]. The appeal of GRADE lies in its ability to provide structure and transparency in the usually complex process of making evidence-based recommendations. It requires a clear clinical question and outcomes important to the patient to be defined from the outset, followed by a structured systematic review of the available evidence. The quality of the evidence is then assessed by considering eight criteria, of which five criteria such as risk of bias, indirectness, inconsistency, imprecision, and publication bias are used to downgrade the quality of evidence. Three other criteria such as magnitude of the effect, dose−response relation in the effect, and opposing plausible residual bias or confounding can be used to upgrade the

This work has been fully funded by the DECIDE Project which is funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 258583.

Competing interests: The authors declare that they have no competing interests.

* Corresponding author. Tel.: +31-20-56-66877; fax: +31-20-69-12683.

E-mail address: g.gopalakrishna@amc.uva.nl (G. Gopalakrishna).

0895-4356
http://dx.doi.org/10.1016/j.jclinepi.2014.01.006

What is new?

Key findings

- When defining the key question(s) for assessing the quality of evidence, a clear distinction is needed between test accuracy and patient-important outcome(s) as the choice outcome. Grading of Recommendations Assessment, Development and Evaluation (GRADE) criteria such as ''inconsistency,'' ''imprecision,'' and ''publication bias'' were challenging to interpret and apply as was the application of the criteria to comparative test accuracy evidence.

What this adds to what was known?

- The current publications on the GRADE for diagnostics approach present an explanation of the approach. In contrast, this article describes the ''practical'' application of the approach when used to rate a body of evidence such as diagnostic tests accuracy review. It outlines a number of real-life challenges and considerations a user of this approach may encounter and provides suggestions on how these can be addressed.

What is the implication and what should change now?

- Explicit guidance and worked examples illustrating the application of the GRADE criteria of inconsistency, imprecision, and publication bias would facilitate the use of the methodology when rating diagnostic test accuracy evidence. Guidance on the translation of a Quality Assessment of Diagnostic Accuracy Studies (QUADAS) 2 risk of bias and applicability assessment to the corresponding GRADE criteria of risk of bias and indirectness would help users in the use of the GRADE approach.

quality of the evidence (Fig. 1). To come to a recommendation based on the available body of evidence, its quality, assessed according to these eight GRADE criteria, is then considered in the context of benefits vs. harms of the test or intervention in question, patients' values and preferences, and resource implications.

Although the GRADE methodology is developed as a generic tool that can be used to assess the quality of evidence for different health care questions, it has most extensively been used for grading the quality of evidence and strength of recommendations for therapeutic questions [3,6,7] and to a lesser degree in the area of medical testing [8–13]. The GRADE approach to grading evidence about the use of medical tests emphasizes the importance of making decisions based on the impact on patient outcomes. It suggests that specific PICO-styled questions about medical tests should specify patients for whom the diagnostic testing is being considered (P), the index test or diagnostic strategy (I), the comparator test or testing strategy (C), and the patient-important outcomes (O) related to the use of the test in question. In this approach, the patient-important outcomes are defined as the desirable and undesirable consequences related to patients being correctly or incorrectly classified as having or not having a given condition, that is, into one of the four test accuracy categories: true positives (TPs), true negatives (TNs), false negatives (FNs), and false positives (FPs). As such, diagnostic test accuracy is considered a surrogate outcome to patient-important outcomes such as morbidity, mortality, or quality of life measures. [9,14,15]. As the evidence base of diagnostic tests is usually restricted to studies that do not measure patient-important outcomes directly, the GRADE for diagnostic approach specifies the need for a clinical judgment making step to link test accuracy to health outcomes.

Research around the conduct and evaluation of test accuracy evidence is increasing [16–18]. However, research evaluating the direct impact of tests on patient outcomes remains sparse for a number of documented reasons [19]. This implies that decision makers often need to consider the potential impact of tests on patient outcomes based solely on the test accuracy evidence. This, together with the increasing uptake of GRADE as a preferred method of evidence appraisal [6], prompted our interest to document issues and challenges a systematic reviewer or guideline panel might counter when applying the GRADE criteria to appraise evidence on test accuracy. We were especially interested in applying the GRADE criteria to evaluate Cochrane diagnostic tests accuracy reviews as both the GRADE Working Group and the Cochrane Collaboration are beginning to work closely together to align the reporting and evaluation of test accuracy evidence.

We focused on the application of those GRADE criteria for which published guidance is currently available. These are the five criteria for downgrading the quality of evidence. We felt that there was not sufficient guidance available yet on the application of the other three GRADE criteria for upgrading evidence related to tests (Fig. 1). The GRADE Working Group is currently preparing two additional publications, which present how the concept of the GRADE approach can be applied to rating the quality of diagnostic test accuracy evidence. Our study is a qualitative evaluation of using the approach in practice, based on current published guidance. It is relevant and useful to ongoing users who need to continue to appraise diagnostic test accuracy evidence based on GRADE for diagnostics guidance that is available in the literature at present.

2. Methods

We assessed the quality of the evidence synthesized in three Cochrane diagnostic test accuracy reviews [20–22].

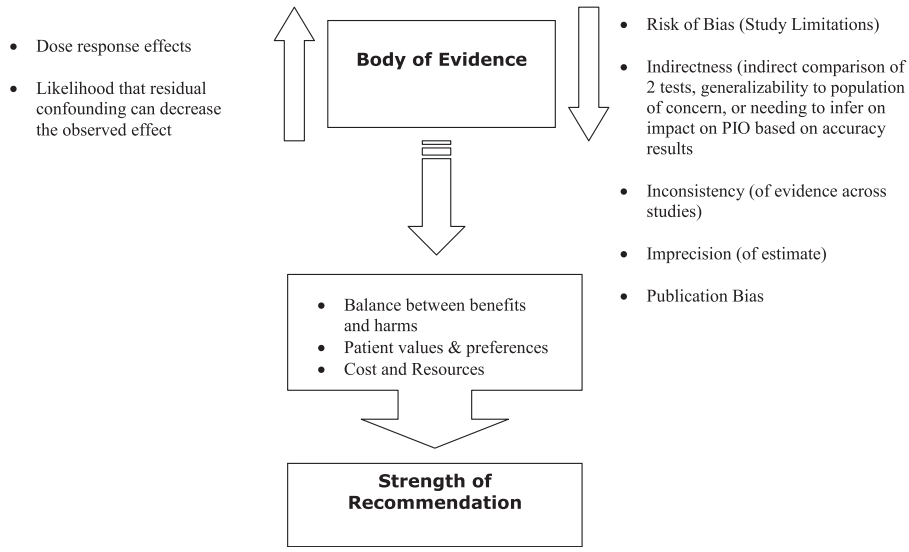

Fig. 1. An overview of GRADE criteria for rating a body of evidence and developing recommendations. GRADE, Grading of Recommendations Assessment, Development and Evaluation.

At the time of the conduct of this study between January and March 2012, there were a total of eight published test accuracy reviews in the Cochrane Library. The field of diagnostic test accuracy review is a relatively new and rapidly evolving field in which the first Cochrane diagnostic test accuracy review appeared just 5 years ago. With this in mind, we made a purposeful selection of reviews that were more recently published while trying to achieve diversity in clinical areas and methodological issues from what was available in the Cochrane Library at that time.

Table 1 provides a brief description of each of the three selected reviews. None of the reviews used the GRADE methodology to evaluate the evidence. All the three reviews used the original Quality Assessment of Diagnostic Accuracy Studies (QUADAS) instrument to assess the included diagnostic accuracy studies [23]. We did not perform any additional formal quality appraisal of each review.

For the review by Abba et al. [22], we chose to evaluate the evidence for the two most frequently reported rapid diagnostic tests (Paracheck and ParaSight). Although the two tests were not directly evaluated against one another in any of the studies included in the review, we decided to rate the quality of the evidence for these two tests as a comparative test accuracy review.

A convenience sample of eight assessors with expertise in the GRADE methodology and/or diagnostic test accuracy reviews were invited to participate in this study. The assessors chosen had varying levels of expertise in the GRADE methodology ranging from proficient to less experienced GRADE users. A protocol was developed and distributed among the assessors. It outlined the publications on GRADE for diagnostics to be used as guidance when applying the GRADE criteria [9,14,15] and a template of a GRADE "evidence profile" which assessors could use. At least three assessors independently rated each review according to the five GRADE criteria for downgrading quality of evidence [14] (Table 1).

Assessors were instructed to explain each judgment that they made about the quality of the evidence and to document all considerations. Two teleconferences were held to share experiences and discuss the challenges encountered. The focus of the discussions was to document how assessors interpreted the GRADE criteria, the issues faced in doing so, and the rationales used to rate the GRADE criteria. We did not record whether assessors downgraded the evidence for each criteria by 1 or 2 points as per the GRADE methodology for downgrading [15].

One author (G.G.) collated all comments discussed during the teleconferences and individual evidence profiles created by each assessor. These were grouped into different categories, based on content, and circulated among all assessors for validation. Differences were discussed among all assessors.

3. Results

The issues encountered could be grouped into four categories: question formulation, each of the five GRADE criteria for downgrading evidence, issues applicable across all GRADE criteria, and issues related to the comparative test accuracy review (Table 2).

3.1. Question formulation

Six of the eight assessors felt that guidance on the formulation of questions for test accuracy reviews was not explicit enough in the GRADE for diagnostics approach. Across all three systematic reviews, assessors formulated different types of key questions; sometimes, no key question was present. For instance, in the review

Table 1. Description of reviews and distribution of assessors across the three Cochrane reviews

Cochrane diagnostic test accuracy review	Description of review	Number of assessors
Optical coherence tomography (OCT) for detection of macular edema in patients with diabetic retinopathy	This review assessed the diagnostic accuracy of OCT for the detection of diabetic macular edema and/or its more severe form of clinically significant macular edema. The review included nine cohort studies. Unit of analyses in the included studies was the individual eye and not the patient.	3 (R.A.M., M.W.L., and M.M.G.L.)
Physical examination for lumbar radiculopathy due to disc herniation in patients with low back pain.	This review assessed tests performed during physical examination (alone or in combination) to identify radiculopathy due to lower lumbar disc herniation as established during imaging or surgery in patients with low back pain and sciatica. The review included 19 studies (16 cohort studies and 3 case–control studies), of which 1 study was conducted in a primary care setting. A variety of physical examination tests were used in the studies with the straight leg raising test or Lasègue test being the most frequent (15 studies). The included studies used different reference standards, with surgical findings or imaging (CT or MRI) being the most frequent ones (nine and six studies, respectively).	6 (G.G., R.A.M., M.W.L., C.D., C.H., and R.J.P.M.S.)
Rapid diagnostic tests (RDTs) for diagnosing uncomplicated *Plasmodium falciparum* malaria in endemic countries.	This review assessed the diagnostic accuracy of immunochromatography-based RDTs for detecting clinical *P. falciparum* malaria (symptoms suggestive of malaria plus *P. falciparum parasitaemia* detectable by microscopy) in persons living in malaria endemic areas who present to ambulatory health care facilities with symptoms of malaria and to identify which types and brands of commercial test best detect clinical *P. falciparum* malaria. The authors included 111 test evaluations from a total of 74 studies, of which 104 test evaluations were in comparison with microscopy, 2 test evaluations were in comparison with PCR-adjusted microscopy, and 5 studies compared RDTs with PCR only. All studies were consecutive patient series.	6 (G.G., R.A.M., J.B., M.W.L., M.M.G.L., and C.D.)

Abbreviations: CT, computed tomography; MRI, magnetic resonance imaging; PCR, polymerase chain reaction.

by Virgili et al., the question prepared by all three assessors contained different elements. The question of one assessor explicitly included a patient-important outcome and the place of the index test in the test–treatment pathway. The second assessor's question was centered on test accuracy and did not contain any patient-important outcomes, although it did explicitly define the place of the index test in the test–treatment pathway. The third assessor's question had no other elements defined except for the index and the reference tests.

3.2. Issues in applying GRADE criteria in a single test review

3.2.1. Risk of bias

In all three reviews, the methodological quality of the primary studies was assessed with the QUADAS tool [23]. The QUADAS results were used to judge the GRADE criterion risk of bias. Assessors found it difficult to make judgments on the quality of the evidence when a QUADAS item was reported as "unclear." This was resolved through discussion among the assessors on the importance of such an item on the overall impact on the quality of the evidence.

Assessors also found it difficult to make an overall judgment on the extent of bias for the corresponding GRADE criterion as they struggled on how to assess the different items in the QUADAS check list to make a final summary statement on the risk of bias for the overall body of evidence.

3.2.2. Indirectness

Assessors reported two observations with the GRADE criterion "indirectness." First, the perspective from which the assessor was looking at the evidence quality influenced the judgment on indirectness. If an assessor was assessing the evidence with patient outcomes as the objective for rating the evidence, the evidence was downgraded by virtue of the fact that evidence relating to test accuracy is indirect evidence for patient-important outcomes. Assessors who assessed the quality of the evidence with test accuracy as the end outcome, however, focused on assessing the extent of indirectness to the intended testing setting among the included studies. The indirectness of the evidence to patient-important outcomes was not considered an issue by these assessors.

Second, some assessors downgraded the quality of the evidence for indirect comparisons if a patient population had been studied that did not match the spectrum of the population for the intended application of the index test. For instance, the aim of the review by Van der Windt et al. [21] was to assess the accuracy of the index test

Table 2. Summary of main issues in the application of the GRADE domains across three Cochrane diagnostic test accuracy reviews

Key issues identified	Observations
Key question formulation	Key question formulation was not an explicit step; guidance on how these could be defined was also not explicit Assessors whose key questions focused on outcomes that were patient important made different judgments on evidence quality compared with assessors whose key questions focused on test accuracy as the outcome
GRADE domains	
Risk of bias (RoB)	Assessors were unclear on how to judge QUADAS items labeled ''unclear''
Indirectness	(1) Issues on applicability of findings to patient population of interest[a] (2) Test accuracy is inherently indirect evidence for patient outcomes, resulting in default downgrading of the quality
Inconsistency	Assessors used different rationales for downgrading (eg, confidence interval overlap, unexplained heterogeneity, inconsistent use of test threshold positivity, and variable reference standard definitions)
Imprecision	Assessors used different rationales for downgrading (eg, small study numbers, wide confidence intervals)
Publication bias	Assessors were unclear on how to assess this
Across all GRADE domains	Reviewers had to be conscious to not double downgrade on a single factor
Additional points for comparative test review[b]	(1) For an indirect comparison of two index tests, the quality of the assessment of test accuracy for each test needed to be assessed first and then the quality of the comparison (2) When making the relative comparison, the score for each GRADE domain (eg, RoB, indirectness, etc.) was determined as the lower of the two scores for that domain for each index test compared with its reference standard (3) The overall quality of evidence (for an indirect comparison of two index tests) was further downgraded by one level for indirectness

Abbreviations: GRADE, Grading of Recommendations Assessment, Development and Evaluation; QUADAS, Quality Assessment of Diagnostic Accuracy Studies.

[a] Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain (Review) [21].

[b] Rapid diagnostic tests for diagnosing uncomplicated P.falciparum malaria in endemic countries (Review) [22].

(elements of physical examinations done singularly or in combination) in the primary care setting. Eight of nine studies in this review, however, were performed in a secondary or tertiary care setting. Hence, the evidence was downgraded for indirectness.

3.2.3. Inconsistency

The two main rational used to downgrade the evidence for this criterion were based on the presence of unexplained heterogeneity among studies and/or the degree of overlap in the confidence intervals (CIs) of the accuracy estimates. If the results of the studies were in the same direction with overlapping CIs, inconsistency was deemed unlikely.

3.2.4. Imprecision

In the review by Virgili et al., the unit of analysis was not the individual subject but the individual eye, which made it challenging to rate the GRADE criterion of imprecision. CIs based on number of eyes tested could give a false sense of a more precise estimate in comparison with an estimate based on the number of subjects tested.

Generally, assessors found it difficult to define how wide a CI should be before it was considered imprecise. Without clinical expertise on the topic of the review or in the absence of guidance in the review itself, assessors tended to downgrade the evidence for imprecision based on individual judgment. In the review by Virgili et al., the 95% CI for specificity was reported as 0.74–0.92 and that for sensitivity as 0.74–0.84. Although all three assessors felt that the CI for specificity represented a wide CI and should be downgraded, the CI for the sensitivity estimate resulted in conflicting views: two assessors felt that the sensitivity estimate did not have a wide CI and did not downgrade the evidence, whereas the third assessor was unsure. One assessor worked out the projected range of subjects for each of the four test accuracy categories (TP, FN, TN, and FP) for a given prevalence of the target condition in question. From there, the assessor made a judgment on imprecision based on the projected numbers for FN and FP derived from this calculation.

3.2.5. Publication bias

Assessors unanimously agreed that guidance was needed on how to appraise test accuracy evidence for publication bias: Seven of the eight assessors did not score the evidence for this criterion because they felt that they had insufficient guidance on its application [24]. One assessor made a judgment based on the assessor's own interpretation of the thoroughness of the search methodology, in terms of databases searched, whether filters had been used, and if grey literature was included.

3.2.6. General comments

In a number of instances, assessors had to be cautious to not double downgrade the evidence for more than one GRADE criterion. For instance, 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.''

3.3. Issues in applying the GRADE criteria in a comparative test review

The review by Abba et al. [22] did not contain any studies that compared both the index tests of interest (ie, Paracheck and Parasight) directly to one another. As a result, assessors felt that they had to make an indirect comparison of the two tests, which raised three observations in the group. Four of the six assessors rated the quality of the evidence for both the index tests separately before proceeding to rate the quality of the evidence for the indirect comparison which was based on the lower of two scores for that same criterion for each index test. Assessors unanimously agreed that the overall quality of evidence for an indirect comparison of two index tests should be further downgraded an additional level because the comparison was not a direct one.

4. Discussion

4.1. Summary of results

When defining the key question(s) for assessing the quality of evidence, a clear distinction is needed between test accuracy and patient-important outcome(s) as the choice outcome. GRADE criteria such as inconsistency, imprecision, and publication bias were challenging to interpret and apply as was the application of the criteria to comparative test accuracy evidence. Appendix (www.jclinepi.com) provides a summary comparison of similarities and differences between the GRADE for interventions and GRADE for diagnostics approaches.

4.2. Areas for further development in the GRADE for diagnostics approach

Most assessors felt that the development of key questions and their elements were not defined explicitly enough in the available GRADE for diagnostics guidance. A reason for this maybe that the ''PICO'' question formulation suggested as guidance for the development of a diagnostic key question may not be directly applicable [25]. Huang et al. demonstrated this when they analyzed 59 real-world clinical questions and showed that one of the limitations of the direct application of the PICO to formulating diagnostic questions was the absence of a clear distinction between population and disease, both elements of which can influence the accuracy of a test.

The other issue our study highlighted was that a clear distinction is needed between test accuracy and patient-important outcome(s) as the outcome included impacts judgments about evidence quality. Guidance would be helpful on developing key questions that are applicable to both types of outcome.

The operationalization of the criteria inconsistency, imprecision, and publication bias would be better facilitated with worked examples as have been provided by the GRADE Working Group for intervention research [26–30]. Such examples will be a part of the next publications on the GRADE for diagnostics approach. In our study, assessors used their own knowledge and discussion among the group to rate the evidence from the reviews for these three criteria. The application of the GRADE criterion inconsistency in intervention evidence states explicitly that unexplained heterogeneity in studies is a clear reason for downgrading. Exploring sources of heterogeneity is often problematic in test accuracy reviews because of small study numbers and poor reporting of included studies [21]. The lack of appropriate statistical methods for assessing study heterogeneity in diagnostic studies further complicates this issue [31] (Appendix at www.jclinepi.com). This means that evidence on tests will often be downgraded for this criterion.

Assessors dealt with rating the evidence quality for imprecision by making judgments on how wide they felt that the CI was for the test accuracy measures of sensitivity and specificity. One assessor's method of calculating the projected range of subjects that would be classified into each of the four accuracy categories may be one way of providing an empirical basis for making judgments on this criterion. Such an approach puts into perspective the projected range of patients that would test as TP, TN, FP, and FN for a given prevalence of the target condition, hence giving an indication of the impact of a test's accuracy in a specific population.

Consider a test whose 95% CI for sensitivity is 50–90%. If the prevalence of the disease is 1% that means the 95% CI of the absolute number of patients diagnosed as TPs works out to be between 5 and 9 for every 1,000 patients tested. The number of patients falsely diagnosed as ''not'' having the disease (FN) will be between 1 and 5 per 1,000 patients tested. One would agree that the 95% CI for TP and FN do not appear to be very large or in this case imprecise at this specified pretest probability.

Now, consider the same test with the same 95% CI range for sensitivity but with a pretest probability of 20%. This now means that 100–180 patients will be correctly diagnosed every 1,000 tested (TP), whereas anything between 20 and 100 patients maybe misdiagnosed as not having the disease (FN). This example illustrates one way in which judgments on imprecision can be made.

This issue of how the GRADE criterion imprecision should be judged is also encountered in intervention research where the same considerations are faced on whether to use relative risk, risk difference, or number needed to treat as the measurement on which imprecision should be judged.

Publication bias was the last criterion that assessors in this study found challenging to apply. This is not unique to diagnostic test evidence as indicated in recent GRADE publications [28]. The lack of consensus methods for assessing publication bias for test accuracy reviews is an added complication to this issue. Furthermore, unlike the increasing awareness and compliance around registration of trials [32] that allow for a way of tracking publication bias, the lack of any kind of a registry for test accuracy

studies in the foreseeable future further compounds this issue of assessing publication bias in test accuracy reviews.

4.3. Considerations for authors of primary studies and systematic reviews of diagnostic test accuracy

As tests are hardly ever conducted in isolation but as part of a test–treatment strategy, it may be helpful for authors conducting primary test accuracy studies and authors of test accuracy reviews to consider defining the test–treatment pathway of the tests being evaluated in their research. This could eventually help not only in selecting appropriate index–reference tests for comparisons but also in defining focused clinical research questions whose findings can have meaningful impact in clinical practice and policy.

This issue was particularly highlighted in the review by van der Windt et al. [21]. In eight of the nine studies in this review, the index test (physical examinations done singularly or in combination) was compared with reference standards (magnetic resonance imaging and surgery) that were relevant to a different patient population and care setting than that of the index test. This resulted in comparisons of test accuracy between different study populations that limited applicability of review findings to primary care where the index test is most commonly used, and therefore, where such evidence synthesis is likely to have the most impact on clinical practice and policy. Van der Windt et al. mention the need for future primary studies to be drawn from the same population in which the index and reference tests would be applied in practice and identify the need for defining a ''diagnostic algorithm.''

This leads us to suggest that the definition of a test–treatment pathway in both diagnostic test accuracy studies and in test accuracy reviews could be a meaningful addition to ensure that we ask clinically relevant questions and design equally relevant studies to answer these questions [33]. Although the Cochrane Collaboration now requires the mandatory inclusion of a clinical pathway in all Cochrane diagnostic test accuracy reviews, there is no explicit guidance on how this should be done or what elements such a pathway should contain.

Several assessors struggled with rating the evidence when the QUADAS [17,23] items were scored as unclear by the review author. The introduction of the QUADAS-2 tool [17] could help address this concern as it calls for users to provide judgments based on the relative importance of individual scoring questions. QUADAS-2 requires an overall risk of bias judgment to be made and separately assesses the applicability of each QUADAS criteria.

Both QUADAS-2 and GRADE for diagnostics are methodological quality assessment tools gaining increasing usage [34]. Although each is meant for a different purpose, QUADAS for the methodological assessment of a diagnostic accuracy study and GRADE for diagnostics, for the evaluation of an aggregated body of evidence, both tools address applicability of the evidence: GRADE as part of its indirectness criterion and QUADAS-2 as part of its first three domains relating to patient selection, index test, and reference standard. Users of both the tools may find it useful to have guidance on how to translate applicability judgments made using the QUADAS-2 tool to judgments about indirectness for GRADE.

Readers of this study should note that the format presented in this study (Table 3) is a work in progress. The GRADE Working Group and the Cochrane Diagnostic Test Accuracy Methods Group are currently collaborating to provide authors of future test accuracy reviews with specific guidance on the format and type of tables to be included in a Cochrane diagnostic test accuracy review.

4.4. Other considerations

Consideration of the test's usefulness, its clinical utility, and the effects on patient outcomes are increasingly acknowledged as being guiding factors when making guidelines for medical tests [35,36]. However, this remains a challenge in medical test guideline development, in which the majority of the evidence base is in the form of accuracy studies.

This raised the issue among the assessors of this study as to whether the indirectness of accuracy studies on patient outcomes warrants a downgrading of the evidence by default. The consequence of this default position would be that the quality of the evidence from test accuracy studies would always be considered of low quality by virtue of the fact that it is indirect evidence for patient outcomes. Users of the GRADE for diagnostics approach may therefore always have only low-quality evidence as long as the type of evidence available on tests is restricted to evidence on test accuracy. However, default downgrading of test accuracy evidence on the basis of indirectness to patient outcomes may not necessarily always be justified. Consider a scenario involving the evaluation of a replacement test shown to have the same or superior accuracy than its comparator test for the diagnosis of a particular condition. In such a scenario, the link to patient outcomes becomes less of an issue and one can accept test accuracy evidence alone as being sufficient without the need for downgrading for indirectness.

Given the challenges assessors in this study found in the interpretation and application of certain GRADE criteria—particularly, those of imprecision, inconsistency and publication bias—guideline development groups which include more than one test accuracy review or involve more than one review team, may find it useful to discuss in advance the application of these criteria so a consistent approach can be applied throughout.

4.5. Limitations of this study

Some of the difficulties faced by this group when making judgments on clinically related issues may not have occurred if topic-specific clinical experts were involved. For instance, assessors found it difficult to determine a clinically relevant sensitivity and specificity value. This made it challenging for

Table 3. Example of an evidence profile created during the study

		Factors that may decrease quality of evidence						Number per 1,000 tested for given prevalence of target condition[1]		
Test result	Study design	Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Test property (95% CI)	Test result	58%	82%	98%
Sensitivity (TP + FN)	Eight historical cohort + one case–control	Serious[2]	Serious[3]	Unclear on how to assess	No	Undetected	0.92 (0.87, 0.95)	TPs	534 (505, 551)	751 (713, 779)	902 (853, 931)
								FNs	46 (29, 75)	66 (41, 107)	78 (49, 127)
Specificity (FP + TN)	Eight historical cohort + one case–control	Serious[2]	Serious[3]	Unclear on how to assess	Serious[4]	Undetected	0.28 (0.18, 0.40)	FPs	284 (252, 344)	130 (108, 148)	14 (12, 16)
								TNs	118 (76, 168)	50 (32, 72)	6 (4, 8)

Abbreviations: CI, confidence interval; TP, true positive; FN, false negative; FP, false positive; TN, true negative.

Cochrane review entitled "Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain" [21]. Population: patients with low back pain; target condition: radiculopathy due to lumbar disc herniation; index test: SLR with variable previous testing and variable criteria for test positivity; reference standard: findings at surgery; setting: primary and secondary care.

[1]Prevalence was based on range in included studies.

[2]Many items are unclear on the QUADAS assessment.

[3]Studies not done in primary care setting. Very high prevalence of condition in included studies.

[4]Very wide CI for specificity.

the assessors to define clinically acceptable sensitivity and specificity thresholds. The involvement of a topic-specific clinical expert would have aided in understanding the impact of the downstream consequences of FNs and FPs, thereby helping to make judgments on which would be the more critical accuracy measures to focus on and what would be acceptable thresholds for FNs and FP numbers.

The reviews chosen for this study were Cochrane diagnostic reviews. Given the generally rigorous methodology of Cochrane reviews, it is likely that additional issues could have arisen if non-Cochrane diagnostic test reviews were included.

We acknowledge that the problems we faced in the key question formulation might be due in part to the fact that we had not provided a specific question for assessors to base their evidence evaluation upon at the start. In a real-world situation, it is likely that a systematic reviewer or guideline developer would have a specific question developed on which they would base their evidence appraisal. That said, we still believe that there is a need to review the applicability of the PICO approach and provide specific guidance on the development of key questions related to medical test evaluation.

5. Conclusions

This study shows that the application of the GRADE criteria for appraising evidence on tests accuracy needs further development before it can be widely applied. An explicit systematic approach to defining the key question, including the test–treatment pathway, and clear guidance for the application of the GRADE criteria of inconsistency, imprecision, and publication bias are needed.

Given the increasing use of the GRADE approach [6] and growing collaboration between GRADE and Cochrane, it maybe useful for future authors of Cochrane test accuracy reviews to consider applying GRADE to their evidence appraisal, particularly on issues related to imprecision, inconsistency, and publication bias which are currently variably addressed by review authors. A more consistent and transparent approach in evaluating test accuracy evidence would help bring better clarity and consistency to such systematic reviews of diagnostic test accuracy which ultimately would also better guide decision makers on to the prudent appraisal of such evidence.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.01.006.

References

[1] Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to

recommendation—determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726−35.
[2] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ 2008;336:995−8.
[3] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.
[4] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.
[5] Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol 2013;66:151−7.
[6] Thornton J, Alderson P, Tan T, Turner C, Latchem S, Shaw E, et al. Introducing GRADE across the NICE clinical guideline program. J Clin Epidemiol 2013;66:124−31.
[7] Treweek S, Oxman AD, Alderson P, Bossuyt PM, Brandt L, Brozek J, et al. Developing and evaluating communication strategies to support informed decisions and practice based on evidence (DECIDE): protocol and preliminary results. Implement Sci 2013;8:6.
[8] World Health Organisation. Policy statement: automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB/RIF system. Geneva, Switzerland: World Health Organisation; 2011.
[9] Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implement Sci 2011;6:62.
[10] Guyatt GH, Akl EA, Crowther M, Schunemann HJ, Gutterman DD, Zelman LS. Introduction to the ninth edition: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141(2 Suppl):48S−52S.
[11] Fiocchi A, Schunemann HJ, Brozek J, Restani P, Beyer K, Troncone R, et al. Diagnosis and Rationale for Action Against Cow's Milk Allergy (DRACMA): a summary report. J Allergy Clin Immunol 2010;126(6):1119−28.
[12] Leung AN, Bull TM, Jaeschke R, Lockwood CJ, Boiselle PM, Hurwitz LM, et al. An official American Thoracic Society/Society of Thoracic Radiology clinical practice guideline: evaluation of suspected pulmonary embolism in pregnancy. Am J Respir Crit Care Med 2011;184:1200−8.
[13] Bates SM, Jaeschke R, Stevens SM, Goodacre S, Wells PS, Stevenson MD, et al. Diagnosis of DVT: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141(2 Suppl):e351S−418S.
[14] Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy 2009;64(8):1109−16.
[15] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106−10.
[16] Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889−97.
[17] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529−36.
[18] 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. BMJ 2003;326:41−4.
[19] Ferrante di RL, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012;344:e686.
[20] Virgili G, Menchini F, Dimastrogiovanni AF, Rapizzi E, Menchini U, Bandello F, et al. Optical coherence tomography versus stereoscopic fundus photography or biomicroscopy for diagnosing diabetic macular edema: a systematic review. Invest Ophthalmol Vis Sci 2007;48:4963−73.
[21] van der Windt DA, Simons E, Riphagen II, Ammendolia C, Verhagen AP, Laslett M, et al. Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. Cochrane Database Syst Rev 2010;(2):CD007431.
[22] Abba K, Deeks JJ, Olliaro P, Naing CM, Jackson SM, Takwoingi Y, et al. Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries. Cochrane Database Syst Rev 2011;(7):CD008122.
[23] 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.
[24] 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.
[25] Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc 2006;359−63.
[26] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines: 6. Rating the quality of evidence—imprecision. J Clin Epidemiol 2011;64:1283−93.
[27] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol 2011;64:1294−302.
[28] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. J Clin Epidemiol 2011;64:1277−82.
[29] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol 2011;64:1303−10.
[30] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.
[31] Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and Presenting Results. Deeks JJ, Bossuyt PM, Gatsonis C. Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 1.0(10). The Cochrane Collaboration; Birmingham, UK 2010.
[32] Laine C, Horton R, Deangelis CD, Drazen JM, Frizelle FA, Godlee F, et al. Clinical trial registration—looking back and moving ahead. N Engl J Med 2007;356:2734−6.
[33] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.
[34] Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. BMC Med Res Methodol 2011;11:27.
[35] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ 2008;336:1049−51.
[36] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089−92.