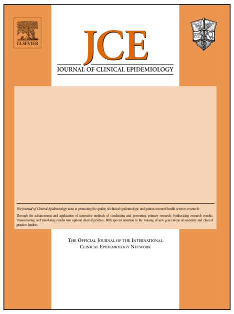


GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy

Holger J. Schünemann, Reem A. Mustafa, Jan Brozek, Karen R. Steingart, Mariska Leeflang, Mohammad Hassan Murad, Patrick Bossuyt, Paul Glasziou, Roman Jaeschke, Stefan Lange, Joerg Meerpohl, Miranda Langendam, Monica Hultcrantz, Gunn E. Vist, Elie A. Akl, Mark Helfand, Nancy Santesso, Lotty Hooft, Rob Scholten, Måns Rosen, Anne Rutjes, Mark Crowther, Paola Muti, Heike Raatz, Mohammed T. Ansari, John Williams, Regina Kunz, Jeff Harris, Ingrid Arévalo Rodriguez, Mikashmi Kohli, Gordon H. Guyatt, for the GRADE Working Group

PII: S0895-4356(19)30673-0

DOI: https://doi.org/10.1016/j.jclinepi.2019.12.020

Reference: JCE 10046

To appear in: *Journal of Clinical Epidemiology*

Received Date: 27 July 2019

Revised Date: 28 November 2019

Accepted Date: 30 December 2019

GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy

Holger J Schünemann[1,2], Reem A. Mustafa[1,3], Jan Brozek[1,2], Karen R Steingart[4], Mariska Leeflang[5], Mohammad Hassan Murad[6], Patrick Bossuyt[5], Paul Glasziou[7], Roman Jaeschke[1,2], Stefan Lange[8], Joerg Meerpohl[9], Miranda Langendam[5], Monica Hultcrantz[10], Gunn E Vist[11], Elie A Akl[12], Mark Helfand[13], Nancy Santesso[1,2], Lotty Hooft[14], Rob Scholten[14], Måns Rosen[10], Anne Rutjes[15], Mark Crowther[1,2], Paola Muti[16], Heike Raatz[17], Mohammed T. Ansari[18], John Williams[19], Regina Kunz[20], Jeff Harris[21], Ingrid Arévalo Rodriguez[22], Mikashmi Kohli[23], Gordon H Guyatt[1,2,3] for the GRADE Working Group

1. Department of Health Research Methods, Evidence, and Impact, McMaster GRADE centre, 1280 Main Street West, McMaster University, Hamilton, Ontario L8S4K1, Canada,
2. Department of Medicine, 1280 Main Street West, McMaster University, Hamilton, Ontario L8S4K1, Canada
3. Department of Medicine, University of Kansas Medical Center, Kansas City, Kansas, USA
4. Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK
5. Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Centers, Room J1b-214, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands
6. Division of Preventive Medicine, Mayo Clinic, 200 1st ST. SW, Rochester, MN, 55902, USA
7. CREBP, Faculty Health Science & Medicine, Bond University, Gold Coast, Qld 4229
8. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen/Institute for Quality and Efficiency in Health Care (IQWiG), Im Mediapark 8, 50670 Köln, Germany Cologne, Germany
9. Institute for Evidence in Medicine, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany & Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany
10. Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), S:t Eriksgatan 117, SE-102 33, Stockholm, Sweden
11. Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway
12. Department of Internal Medicine, American University of Beirut, Riad-El-Solh Beirut, Beirut 1107 2020, Lebanon.
13. Oregon Evidence-based Practice Center, Oregon Health & Science University, Portland VA Medical Center, Portland, Oregon
14. Cochrane Netherlands/Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508GA Utrecht, The Netherlands
15. Clinical Trial Unit (CTU) Bern, Institute of Primary Health Care; Institute of Social and Preventive Medicine, University of Bern, Switzerland
16. Department of Oncology, McMaster University, 711 Concession Street, Hamilton, ON L8V1C3, Canada
17. University of Basel, Klingelbergstrasse 61, CH-4056 Basel, Switzerland & Kleijnen Systematic Reviews Ltd., 6 Escrick Business Park, Escrick, York YO19 6FD, UK

18. School of Epidemiology and Public Health, Faculty of Medicine, Ottawa, Canada
19. Duke University Medical Center and Durham Veterans Affairs Center for Health Services Research in Primary Care Durham, NC 27705, USA
20. Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, Basel, 4031, Switzerland
21. Harris Associates, 386 Richardson Way, Mill Valley, CA 94941, USA
22. Clinical Biostatistics Unit, Ramón y Cajal Hospital (IRYCIS), Madrid, Spain and Division of Research, Fundación Universitaria de Ciencias de la Salud, Hospital de San José/ Hospital Infantil de San José, Bogotá, Colombia.
23. Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1650 Cedar Ave, Montreal, QC, H3G 1A4, Canada

Word count: 3715

Tables: 4

Figures: 3

Key findings

Rating the certainty of the body of evidence (quality of evidence or confidence in estimates) of test accuracy studies differs conceptually but shares the fundamental logic for the domains risk of bias and indirectness of the GRADE approach for intervention, prognostic or other studies.

What this adds to what is known?

Questions about the relative merit of alternative testing strategies in clinical and public health require framing in terms of health outcomes. Evidence evaluation will often, however, begin with an evidence synthesis - ideally a systematic review or health technology assessment - and rating of test accuracy, and subsequently move to evaluation of the evidence linking test accuracy to patient-important and population outcomes. We describe examples for how GRADE has been applied to test accuracy studies in Cochrane and other reviews and World Health Organization and other guidelines focusing on risk of bias and indirectness in part 1 of this article.

What are the implications, what should change now?

Investigators interested in using the GRADE for diagnostic and other healthcare related tests should consider the guidance offered in this article about how to evaluate research that focuses on the impact of tests, specifically from test accuracy studies in the context of risk of bias and indirectness. We provide examples for how to separate indirectness on a population, test intervention, test comparison and outcome levels.

Abstract

Objectives: This article provides updated GRADE guidance about how authors of systematic reviews and health technology assessments (HTA) and guideline developers can assess the results and the certainty of evidence (also known as quality of the evidence or confidence in the estimates) of a body of evidence addressing test accuracy (TA).

Study Design and Setting: We present an overview of the GRADE approach and guidance for rating certainty in TA in clinical and public health and review the presentation of results of a body of evidence regarding tests. Part 1 of the two parts in this 21st guidance article about how to apply GRADE focuses on understanding study design issues in test accuracy, provide an overview of the domains and describe risk of bias and indirectness specifically.

Results: Supplemented by practical examples, we describe how raters of the evidence using GRADE can evaluate study designs focusing on tests and how they apply the GRADE domains risk of bias and indirectness to a body of evidence of TA studies.

Conclusions: Rating the certainty of a body of evidence using GRADE in Cochrane and other reviews and World Health Organization and other guidelines dealing with in TA studies helped refining our approach. The resulting guidance will help applying GRADE successfully for questions and recommendations focusing on tests.

Key words: GRADE, diagnosis, tests, test accuracy, certainty of evidence, diagnostic accuracy

GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy

1.0 Introduction

Previous GRADE articles described the reasons for decreasing and increasing the certainty of a body of evidence; how to perform and present an overall rating of the evidence; how to use evidence to move to decisions and recommendations; guidance for addressing missing outcome data and multiple intervention comparisons; rating evidence regarding values and preferences and the use of GRADE in the context of environmental and public health questions and rapid guidance.(1-21) Clinicians and policy-makers also face choices regarding diagnostic, monitoring or screening tests or test strategies, choices that if made optimally will result in net benefit for people or patient-important outcomes and overall net desirable consequences.(9-11)

However, questions related to tests present unique challenges. Here, we describe how authors of systematic reviews and health technology assessments (HTAs) and guideline developers using GRADE can address the certainty (in this series also referred to as quality or confidence) in a body of evidence from test accuracy (TA) studies, and present the results of their assessment. These articles supplement our previous work addressing GRADE for tests.(11, 22, 23) Part 1 of the two parts in this 21st guidance article is about how to apply GRADE focuses on understanding study design issues in TA, provide an overview of the domains and describe risk of bias and indirectness specifically. Part 2 focuses on the domains imprecision, inconsistency, publication bias, considerations about upgrading and guidance about how to present this information in GRADE summary of findings (SoF) tables .(24)

2.0 Establishing the purpose of a test

Health care providers use tests that are often referred to as “diagnostic” – including signs and symptoms, imaging, biochemistry, psychological assessments, pathology, microbiology and other tests – to guide management and for other purposes.(25) Guideline groups and authors of systematic reviews or HTAs addressing tests must define the purpose and role of the tests in their particular context. This process should begin with determining the existing test or diagnostic pathway – or pathways – for the patient presentation and identify the adverse outcomes that will arise from direct complications of invasive tests or the consequences of true positive, e.g. overdiagnosis, true negatives, false positive and false negative test results. Knowing these adverse outcomes may suggest if an alternative test or strategy that is less invasive or with superior diagnostic properties may result in greater net health benefit which we discuss in more detail in section 2.3. and elsewhere.(26)

The purpose of a test under consideration may be for screening, risk assessment, diagnosis, prognosis, staging, monitoring, or surveillance. The role of a test may be for (i) replacement (i.e., with tests with less burden, invasiveness, cost, or superior accuracy), (ii), triage (i.e., to minimize use of an invasive or expensive test), (iii) add-on (i.e., to further enhance diagnostic accuracy beyond the existing diagnostic pathway) or (iv) parallel or combined testing (i.e. tests that health professionals order and evaluate together to inform a particular diagnosis, table 1).(27-29)

2.1 Positive and negative tests, single tests and test strategies

While some tests report positive and negative results (e.g., pregnancy, HIV infection), other tests report their results as ordinal (e.g., Glasgow coma scale, mini-mental status examination) or continuous variables (e.g., serum ferritin, troponin, hemoglobin), with increasing likelihood of disease or adverse health effects as the test results become more extreme. For simplicity, in this discussion we generally assume an approach that ultimately categorizes test results as positive or negative, in part to describe presence or absence of a target condition. This also recognizes that many tests ultimately lead to treat or do not treat decisions based on the “positive or negative” result of the test.

Clinicians, public health workers and researchers often administer tests as a strategy or package composed of several tests. For example, in managing patients with a diagnosis of cervical intraepithelial neoplasia - a precursor of cervical cancer - based on visual inspection with acetic acid (VIA), clinicians may proceed to treatment directly or further test for human papilloma virus (HPV) to increase the probability of neoplasia being present.(30, 31) A testing strategy may also use an initial sensitive but non-specific test which, if positive, is followed by a more specific test (e.g., testing for HIV includes the use of an ELISA test followed by quantitative HIV RNA determination for those with positive ELISA test results). Thus, one can often think of evaluating or recommending a test strategy rather than a single test, and usually it is a recommendation based on a comparison to alternative test strategies.

2.2. Clear healthcare and clinical questions

Clearly establishing purpose and role of a test or test strategy will lead to the identification of sensible healthcare questions that, similar to other management problems, have four components: population, intervention tests (or strategies), comparison test (or strategy), and the outcomes of interest.(32, 33) Labeling testing as an intervention recognizes the fundamental principle that test results will lead to specific management decisions and those decisions will influence outcomes. Box 1 shows three examples of questions about the use of tests to which we will refer in this and other articles in this series.(11, 26)

Box 1. Examples of questions about tests

Example 1: ***In women at risk for cervical intraepithelial neoplasia (CIN) in low and middle-income settings, what is the impact of testing for presence of human papilloma virus (HPV) instead of VIA on patient and population important outcomes?(30)***

Population: women at risk of cervical cancer in low and middle-income countries

Role: replacement test

Setting: clinics in low and middle income countries

Intervention: one-time screening with HPV and treatment for cervical intraepithelial neoplasia

Comparison: VIA and treatment for cervical intraepithelial neoplasia

Purpose and role of test: diagnosis and replacement of no testing

Outcomes: death from cervical cancer, cervical cancer incidence, CIN recurrence, major bleeding, premature delivery, infertility, major and minor infections, unnecessary treatment and burden, cervical cancer detection during screening

Example 2 (short form focusing on patient outcomes): ***In patients suspected of cow's milk allergy (CMA), what is the impact of skin prick tests versus an oral food challenge with cow's milk on mortality from allergic reactions, allergic reactions, development of other allergies.(34)***

Participants: patients suspected of CMA

Role: replacement test

Setting: specialized clinics

Index (new) test (intervention): IgE skin prick test

Reference test (comparison): no IgE skin prick test

Outcome: test accuracy with health outcome descriptors for the test positives and negatives

Example 3 (test accuracy focused): ***In patients presumed to have tuberculous (TB) meningitis, what is the accuracy of Xpert – a nucleic acid amplification test (NAAT) – for the diagnosis of TB meningitis?***

Participants: patients suspected of having TB meningitis

Prior testing: patients who received Xpert testing may first have undergone a health examination (history and physical examination) and possibly a chest radiograph

Role: replacement test for usual practice

Settings: primarily tertiary care centres (the index test was run in reference laboratories)

Index (new) test (intervention): Xpert

Reference test (comparison): culture

Outcome: test accuracy

The example of using *HPV instead of VIA for screening for cervical precancerous lesions* in Box 1 illustrates one common rationale for a new test: test replacement to avoid a slightly more invasive alternative for a condition amenable to effective treatment.(27) Such a new test would only need to be as accurate as the existing test to demonstrate greater net benefit (by lowering burden and harm). This assumes that the new test similarly categorizes patients at the same stage of disease and that the consequences of the test result, i.e.

management decisions and outcomes, are similar. However, these scenarios are not common.

2.3. Estimating impact on people or patient-important outcomes

Recommendations regarding use of tests require inferences about the consequences of falsely identifying patients as having or not having the disease, but also consequences of correct test results that do not lead to net benefit (e.g. overdiagnosis or lack of treatment effects). If a test fails to improve people-important outcomes (in the context of population or public health) or patient-important outcomes (in the context of clinical care) there is no reason to use it, whatever its accuracy. We will refer to the consequences of alternative testing strategies on these outcomes as “test impact”.

Dealing first with TA, ascertainment of TA relies on the presence of a gold, reference or criterion standard that is used to establish if the target disease is present or absent. Often, an error-free gold standard is unavailable. Moreover, constructs of disease may change (e.g, in oncology, with a superior molecular understanding of the underlying pathologies, or Alzheimer’s dementia). We will use the term gold standard here as representing the “perfect” approach to defining or diagnosing the disease or condition of interest, even if the approach is theoretical or hypothetical. We will use the term “reference standard” for the test or test strategy that is the current best and accepted approach to making a diagnosis against which a comparison with an “index test” (the test under consideration) may be made. A reference standard can consist of the gold standard, but more likely represents a less than ideal reference standard – in which case one could compare the accuracy of the reference standard to the gold standard, if the gold standard is feasible to perform.

However, by definition accuracy cannot be superior to the gold standard. Also by definition, acceptance of a new gold standard, for instance as a result of scientific development, requires consensus, proof of added benefit and acceptance rather than only demonstration of better TA.

Given the importance of focusing on outcomes that are important to people and the uncertainties related to reference and gold standards and the relation between tests and patient or population consequences, the best way to assess a test strategy is a test-treat randomized controlled trial design in which investigators allocate patients to experimental or control testing approaches and measure mortality, morbidity, symptoms, quality of life and resource use.(35)

Figure 1a describes the fundamental elements of study designs of test accuracy studies and the ideal test impact study – that is, a randomized trial of alternative testing and management strategies. Various randomized designs leading to high confidence in estimates exist, including interaction designs that help to directly determine the impact, both positive and negative, of a test on health outcomes.(28, 36, 37) When test impact studies – ideally RCTs but also observational studies – comparing alternative test strategies with direct assessment of patient-important outcomes are available (Figure 1a), guideline panels can use the GRADE approach for other interventions described in prior GRADE articles.(1, 38)

All too frequently, however, management decisions depend on evidence obtained in separate steps. Figure 1b illustrates a generic study structure that guideline developers often have to use to evaluate the impact of testing.(22)

Insert Figure 1 (a) & (b) approximately here

This latter approach links evidence by bringing together TA estimates with evidence regarding subsequent management and the treatment effects associated with that management to model the impact of TA results on health outcomes.(22, 39) In that situation, TA may be considered a surrogate outcome for people-important benefits and harms.(25) Those developing recommendations must make these inferences, and the underlying assumptions about the evidence on which the inferences are based, transparent.

The key questions when using TA as a surrogate are: (i) what outcomes can those with positive and negative test results expect based on the likely subsequent management?; (ii) what will be the relative impact of the testing strategies under consideration on the number of false negatives (people with the disease who are missed) or false positives (people without the disease who are incorrectly considered as having the disease)?; and (iii) how similar or different are people to whom the test is applied in practice (and classified by the alternative testing strategies) to those evaluated in TA studies? An alternative is to abjure making explicit inferences, to provide guidance solely on the basis of TA information and point out that direct evidence for a people-important benefit is lacking – perhaps with a recommendation for generation of such evidence.(40)

2.4. Indirect evidence and impact on patient-important outcomes

Consequences of tests typically go beyond the benefits and harms that one usually considers in assessing therapeutic interventions. We previously described the issues with highly accurate genetic testing for Huntington's chorea, a condition that currently cannot be cured, may provide either welcome reassurance that a patient will not suffer from the condition or the ability to plan for the future knowing that the patient may sadly fall victim.(22) In this case, the ability to plan the future is analogous to the usual outcomes of benefit (e.g. reducing mortality) and harm (e.g. adverse effects), and the benefits of planning require trading off against the downsides of receiving an early diagnosis. In such instances - as in most others - guideline panels would review the evidence and they might find that the evidence does not equivocally support testing, i.e. providing net benefit, because differences in values and preferences are likely to play an important role in making this decision.(41-43)

Thus, inferring from accuracy data that a diagnostic test or strategy improves patient-important outcomes usually requires access to effective management and the values that relate to the anticipated outcomes.(25) In GRADE guidance 22 in this series, we will discuss these issues in greater detail. Now we will focus on the assessment of certainty of evidence TA.

2.5 Judgment about the certainty of the underlying evidence

We will use the systematic review by Kohli et al. (44) to demonstrate how judgements are made (an online supplement provides additional examples). This review looked at Xpert® MTB/RIF (Xpert), a rapid, automated, nucleic acid amplification assay that is widely used for simultaneous detection of *Mycobacterium tuberculosis* complex and rifampicin resistance in sputum specimens. (44) Our focus is on evidence regarding the usefulness of the test in the diagnosis of tuberculous meningitis (Example 3, Box 1).

3.0 Certainty of the evidence from TA studies

In the GRADE approach, appropriately designed TAs (see below) start as high certainty evidence. However, in the context of providing evidence to support guideline recommendations or decisions focusing on these studies in isolation will usually result in low or very low certainty supporting the decision due to indirectness of evidence because TA is a surrogate for the impact of testing on patient-important outcomes. That is, if those developing recommendations or making decisions do not identify and assess the linked evidence, they should rate down the evidence supporting a decision or specifically describing that they only considered the certainty of the TA studies (see article 22 in this series (26)), Table 2 lists factors that influence the certainty of a body of evidence from TA studies.

Insert Table 2 approximately here

In the tuberculous meningitis example, all studies were cross-sectional studies appropriately designed to evaluate test accuracy.(44) The initial rating for the body of evidence for test accuracy studies is high.

3.1. Certainty of the evidence for TA - risk of bias (limitations in the detailed study design and execution)

Researchers have developed several instruments for the evaluation of risk of bias in TA studies.(45-47) For example, a selection of the items of the QUADAS-2 instrument allow transparent assessment of risk of bias based on the features shown in Table 3. (48)

Insert Table 3 approximately here

Appropriate TA studies include patients with an uncertain diagnosis who are representative of the target population. Such studies should preferentially enrol consecutive or randomly selected patients in whom diagnostic uncertainty exists – that is, the sort of patients to whom clinicians would apply the test in the course of regular clinical practice. If studies fail this criterion – and for example enrol severely affected patients and healthy controls – the apparent accuracy of a test is likely to be misleadingly high.(49, 50)

3.1.1. Examples of risk of bias (limitations in the detailed study design and execution) judgments

Appropriate TA studies also involve a comparison between one or more tests under consideration, where all tests are measured against the same reference standard. Investigators' failure to apply the reference standard in all patients increases the risk of bias. The risk of bias may be higher if a composite reference standard is used and the included studies use different ways of ascertaining the presence or absence of a target condition. The risk of bias is further increased if those who conduct or interpret the index test are aware of the results of the reference standard or vice versa.

In our example of Xpert for the diagnosis of tuberculous meningitis, using QUADAS-2, risk of bias was low for the patient selection, index test, and flow and timing domains. For the reference standard domain, of 29 total studies, four (14%) studies had unclear risk of bias. In these four studies, specimens underwent decontamination and it was unclear whether this process affected the reference standard. However, since most studies had low risk of bias, the authors did not downgrade (Figure 2).(44) In another systematic review, using the original QUADAS instrument (QUADAS-2 was introduced in 2011), Steingart and colleagues evaluated the risk of bias of studies of commercial serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis.(51) Most of the 67 included studies did not recruit participants in a random or consecutive manner, and only approximately 50 studies reported blinded interpretation of the serological test result. The authors, therefore, downgraded the certainty of the evidence for risk of bias.

3.2. Certainty of the evidence – applicability and indirectness

Direct evidence comes from research that closely addresses the population we are interested in, compares the interventions in which we are interested and measures the outcomes important to people or patients. Judging indirectness - synonymous with applicability, transferability, generalizability, translatability and external validity of the evidence - includes indirectness related to the downstream consequences. As with therapeutic interventions, indirectness must be assessed in relation to the population (and setting), the intervention (the new or index test), the comparator (another test), and outcomes. We will deal with indirectness related to test outcomes and their impact in more detail in article 22 in this series.(26)

3.2.1. Population indirectness

The chosen patient sample may cause indirectness. Studies may also provide indirect evidence if the target condition of the population is not the same in the studies compared to the question asked and an interaction between the population and test performance is expected. Population indirectness does not only relate to the disease spectrum in the included patients, but also to the setting in which the research was done, prior testing done on the patients, or possible referral paths. For instance, a judgment of indirectness of the population can result from using a different test setting (e.g. patients seen in an emergency department may differ from patients seen in a general practitioner office). Another example may be when guidance is needed for testing in children and the available evidence is based upon studies with adults, or mixed age populations. Prevalence or pretest probability may be a guide to judge whether there is indirectness in the population: is the average prevalence in the available evidence comparable to what is found in practice? Investigators can explore the influence of all these sources of variability in sensitivity analyses.

3.2.2. Intervention(s) or Index test(s) indirectness

Indirectness in the intervention or index test domain may occur when, for example, tests in the studies or reviews found have been implemented with slightly different standards than the standards used in practice, e.g. a country, specialty or health plan, for which the guidance is intended. Different cut-off values or thresholds between settings may lead to indirectness (and often explain inconsistency in sensitivity analyses). Different settings may also introduce intervention indirectness if the test is applied in an emergency department as opposed to a primary care setting (e.g., due to specimen transport or personnel qualifications in emergency departments as opposed to a general practitioner's office). Specimen transfer issues become particularly salient if mechanistic studies have demonstrated that transport conditions affect test performance (e.g. transfer at room temperature versus on ice may induce changes in a serum level of a biomarker). This type of indirectness incorporates concepts of technical variability and test-retest and operator reliability.

3.2.3. Comparator(s) indirectness

Any question about tests will have a comparison at its basis. Unfortunately, accuracy studies often focus on one test only. For example, in a question focusing on the accuracy of D-dimer for pulmonary embolism the comparison may be between different D-dimer tests (TA at different D-dimer levels), between clinical signs and symptoms (standard care; no testing) and no D-dimer (does d-dimer have a higher TA than signs and symptoms alone), or between another test and D-dimer. That means that in case of an explicit comparison with another test or in the key question, this comparator will almost always cause some degree of indirectness in the comparator domain. Also, if comparative accuracy studies compared our test of interest with another test that we are not interested in as our comparator, then that may lead to indirectness. Second, if the clinical question is about the choice between two tests, neither of which is a reference standard, the two tests may be compared directly against the reference test in the same study. For instance, in the example of comparing HPV with VIA all studies independently utilized both tests and compared it against a reference standard (colposcopy with biopsy) to evaluate the incremental or relative accuracy, sometimes expressed as difference in accuracy or comparative accuracy (Figure 3).(31) Alternatively, one might make an indirect comparison based on separate studies in which each test was compared against the reference standard and usually rate down for indirect comparisons.

3.2.4. Outcomes and outcome measures

TA as an outcome will nearly always cause indirectness for guideline developers and other decision makers because the recommendations and decisions should be based upon intervention outcomes that follow from the test results when the available evidence often only includes accuracy as an outcome.

If the key question focuses on diagnostic TA and authors of a systematic review rate the certainty in accuracy outcomes, then there may be no indirectness between the available evidence (accuracy studies) and the desirable outcomes (accuracy measures). The indirectness related to people-important outcomes should then be assessed by those using the review for recommendations or decisions (see GRADE guidance on evidence to decision frameworks and linked evidence (11, 26)). However, there may still be indirectness in the outcome itself, for example because of a different definition of the condition of interest in studies compared to the condition in the systematic review author's, HTA author's or guideline developer's question of interest.

3.2.5. Examples for indirectness

In the tuberculous meningitis example, the three included studies engendered concern about indirectness because patients were evaluated as inpatients in tertiary care centres but tuberculous meningitis is a medical emergency often treated in primary and secondary care hospitals. The raters judged the patients in tertiary care centers to be similar to those in secondary and primary care settings and did not rate down for indirectness. For the index and reference test domains, the authors considered most studies to have no serious concern because both the index test was performed as in routine clinical practice. With regards to the outcomes, the raters did not evaluate the indirectness related to the downstream health consequences in the review and this was left to the decision makers, which we deal with in the next article in this series.(26) Table 4 demonstrates how raters can structure their judgments transparently.

Insert Table 4 approximately here

5.0 Conclusion

The GRADE approach to rating the certainty of evidence for TA is comprehensive and transparent. We have presented an overview of this approach and operatonialized the first two of the GRADE domains, risk of bias and indirectness, to rate the certainty in a body of evidence for TA studies. They have been applied in many systematic reviews and guidelines. In part 2 of this GRADE guidance about rating the certainty of evidence in tests, we will focus on the domains inconsistency, imprecision, rating up and presentation of findings related to TA.(24) In GRADE guidance 22 in this series we will describe how the information from test accuracy can inform the development of recommendations, based on the recognition that test results can be surrogate markers for patient important outcomes.(26)

Disclosure Statement

The authors are members of the GRADE Working Group. ML is Co-convenor of Cochrane's Screening and Diagnostic Test Methods Group.

Acknowledgment
This work was partially funded by a "The human factor, mobility and Marie Curie Actions Scientist Reintegration" European Commission Grant: IGR 42192 – "GRADE" to Dr. Schünemann. We would like to thank the many individuals and organizations who have contributed to the progress of the GRADE approach through hosting of meetings and feedback on the work described in this article. The work on this article has in part been a result of collaborative effort over more than 10 years. We would like to thank Drs. Andrew Oxman, Rob Scholten and Jon J Deeks who participated in conversations and group meetings leading to this approach.

References

1. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.

2. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knotterus A. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2010.

3. Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014;349:g5630.

4. Schunemann HJ, Best D, Vist G, Oxman AD, Group GW. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ. 2003;169(7):677-80.

5. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. ACP J Club. 2008;149(6):2.

6. Spencer FA, Iorio A, You J, Murad MH, Schunemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. Bmj. 2012;345:e7401.

7. 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(7653):1106-10.

8. 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(2):151-7.

9. Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. Bmj. 2016;353:i2089.

10. Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, 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.

11. Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98.

12. Burford BJ, Rehfuess E, Schunemann HJ, Akl EA, Waters E, Armstrong R, et al. Assessing evidence in public health: the added value of GRADE. J Public Health (Oxf). 2012;34(4):631-5.

13. Thayer KA, Schunemann HJ. Using GRADE to respond to health questions with different levels of urgency. Environ Int. 2016;92-93:585-9.

14. Schunemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, et al. Transparent development of the WHO rapid advice guidelines. PLoS Med. 2007;4(5):e119.

15. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. Bmj. 2015;350:h870.

16. Guyatt GH, Ebrahim S, Alonso-Coello P, Johnston BC, Mathioudakis AG, Briel M, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. J Clin Epidemiol. 2017.

17. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, et al. GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int. 2016.

18. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.

19. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;64 (4):380-2.

20. Zhang Y, Alonso Coello P, Guyatt G, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE Guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences - Inconsistency, Imprecision, and other Domains. J Clin Epidemiol. 2018.

21. Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. J Clin Epidemiol. 2018.

22. 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 (Clinical research ed). 2008;336 (7653):1106-10.

23. 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.

24. Schunemann HJ, Mustafa RA, Brozek J, Steingart K, Leeflang M, Murad HM, et al. GRADE guidelines: 21 part 2. Inconsistency, Imprecision, publication bias and other domains for rating the certainty of evidence for test accuracy and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. in press.

25. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. Bmj. 2001;323(7305):157-62.

26. Schunemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82.

27. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92.

28. Mustafa RA, Wiercioch W, Cheung A, Prediger B, Brozek J, Bossuyt P, et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 2: a review of methodological and practical challenges. J Clin Epidemiol. 2017;92:18-28.

29. Schunemann HJ, Mustafa RA. Decision making about healthcare-related tests and diagnostic test strategies. Paper 1: a new series on testing to improve people's health. J Clin Epidemiol. 2017;92:16-7.

30. Santesso N, Mustafa RA, Schunemann HJ, Arbyn M, Blumenthal PD, Cain J, et al. World Health Organization Guidelines for treatment of cervical intraepithelial neoplasia 2-3 and screen-and-treat strategies to prevent cervical cancer. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics. 2016;132(3):252-8.

31. Santesso N, Mustafa RA, Wiercioch W, Kehar R, Gandhi S, Chen Y, et al. Systematic reviews and meta-analyses of benefits and harms of cryotherapy, LEEP, and cold knife

conization to treat cervical intraepithelial neoplasia. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics. 2016;132(3):266-71.
32. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Cmaj. 1988;138(8):697-703.
33. Mulrow C, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. Journal of General Internal Medicine. 1989;4(4):288-95.
34. Fiocchi A, Brozek J, Schunemann H, Bahna SL, von Berg A, Beyer K, et al. World Allergy Organization (WAO) Diagnosis and Rationale for Action against Cow's Milk Allergy (DRACMA) Guidelines. Pediatr Allergy Immunol. 2010;21 Suppl 21:1-125.
35. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet. 2000;356(9244):1844-7.
36. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2005;23(9):2020-7.
37. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol. 2009;62(4):364-73.
38. 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(4):395-400.
39. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Annals of internal medicine. 2006;144(11):850-5.
40. Evans WK, Laupacis A, Gulenchyn KY, Levin L, Levine M. Evidence-based approach to the introduction of positron emission tomography in ontario, Canada. J Clin Oncol. 2009;27(33):5607-13.
41. Maat-Kievit A, Vlis MV-vd, Zoeteweij M, Losekoot M, van Haeringen A, Roos R. Paradox of a better test for Huntington's disease. J Neurol Neurosurg Psychiatry. 2000;69(5):579-83.
42. Walker FMD. Huntington's Disease. Semin Neurol. 2007(02):143-50.
43. Almqvist EW, Brinkman RR, Wiggins S, Hayden MR. Psychological consequences and predictors of adverse events in the first 5 years after predictive testing for Huntington's disease. Clinical Genetics. 2003;64(4):300-9.
44. Kohli M, Schiller I, Dendukuri N, Dheda K, Denkinger CM, Schumacher SG, et al. Xpert((R)) MTB/RIF assay for extrapulmonary tuberculosis and rifampicin resistance. Cochrane Database Syst Rev. 2018;8:CD012768.
45. 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. Annals of internal medicine. 2003;138(1):40-4.
46. 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.
47. 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.

48. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of internal medicine. 2011;155(8):529-U104.
49. 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.
50. 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.
51. Steingart KR, Flores LL, Dendukuri N, Schiller I, Laal S, Ramsay A, et al. Commercial Serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis: An updated systematic review and Meta-Analysis. PLoS Medicine. 2011;8 (8)(e1001062).
52. 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.
53. Moe G, Howlett J, Januzzi JL, Zowall H. Canadian Multicenter Improved Management of Patients With Congestive Heart Failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian Prospective Randomized Multicenter IMPROVE-CHF study. Circulation. 2007;115(24):3103-10.
54. 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.
55. Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? Canadian Association of Radiologists journal = Journal l'Association canadienne des radiologistes. 2002;53(3):144-8.

Table 1. Possible roles of new diagnostic tests (27, 28)

Replacement	A new test might substitute an old one, because it is more accurate, less invasive, less risky or uncomfortable for patients, organizationally or technically less challenging, quicker to yield results or more easily interpreted, or less costly.
Triage	A new test is added before the existing diagnostic pathway and only patients with a particular result on the triage test continue the testing pathway; triage tests are not necessarily more accurate but usually simpler and less costly.
Add-on	A new test is added after the existing diagnostic pathway and may be used to limit the number of either false positive or false negative results after the existing diagnostic pathway; add-on tests are usually more accurate but otherwise less attractive than existing tests.
Parallel or combined	A new test which is intended to be used concurrently with an existing test. Both, the results of the existing and the parallel test are utilized for making a diagnosis and to determine management.

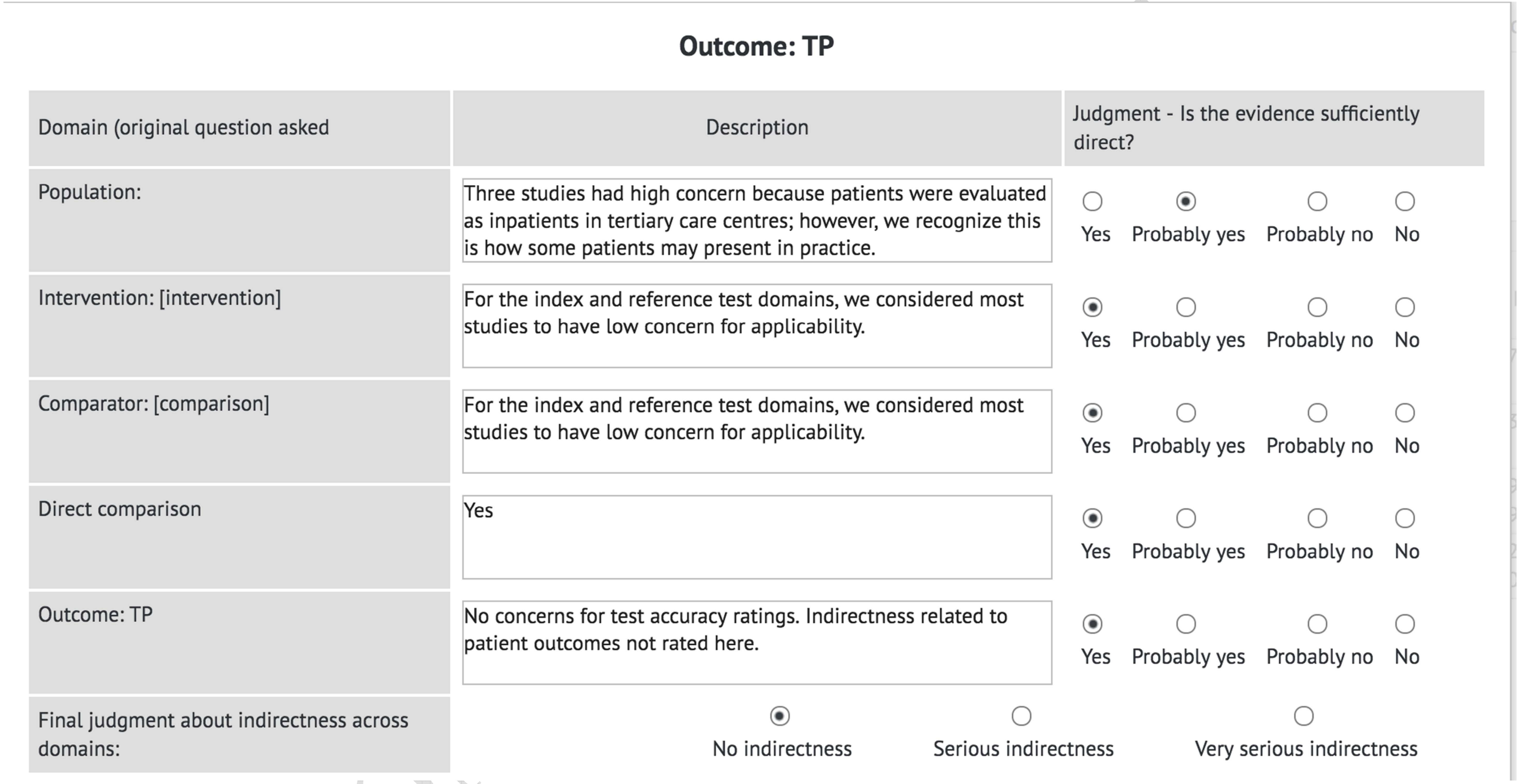
Table 2. Factors that decrease the certainty of evidence for studies of test accuracy and how they differ from evidence for other interventions

Factors that determine and can decrease the certainty of evidence	**Explanations and how the factor may differ from the certainty of evidence for other interventions**
Study design	Different criteria for accuracy studies Cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard (best possible alternative test strategy) start as high certainty but may be rated down to moderate, low or very low depending on other factors.
Risk of bias (limitations in study design and execution)	Different criteria for accuracy studies • Representativeness of the population that was intended to be sampled. • Independent comparison with the best alternative test strategy. • All enrolled patients should receive the new test and the best alternative test strategy. • Diagnostic uncertainty should be given. • Is the reference standard likely to correctly classify the target condition? • QADAS2 is an acceptable tool.
Indirectness and applicability Patient population, index test, comparison test, indirect comparisons of tests and indirect outcomes	Similar criteria to therapy questions The certainty of evidence can decrease if there are important differences between the populations studied and those for whom the recommendation is intended (in prior testing, the spectrum of disease or co-morbidity); if there are important differences in the tests studied and the diagnostic expertise of those applying them in the studies compared to the settings for which the recommendations are intended; or if the tests being compared are each compared to a reference (gold) standard in different studies and not directly compared in the same studies. Similar criteria to therapy questions Guideline groups assessing diagnostic tests often face an absence of direct evidence about impact on patient-important outcomes. They must make deductions from diagnostic test studies about the balance between the presumed influences on patient-important outcomes of any differences in true and false positives and true and false negatives in relationship to test complications and costs. Therefore, accuracy studies typically provide low quality evidence for making recommendations due to indirectness of the outcomes, similar to surrogate outcomes for treatments. Guideline groups should therefore identify linke evidence that informs about the consequences of the accuracy outcomes (26)
Important Inconsistency in study results	Similar criteria to therapy questions For accuracy studies unexplained inconsistency in sensitivity, specificity or likelihood ratios (rather than relative risks or mean differences) can lower the certainty of evidence.
Imprecise evidence	Similar criteria to therapy questions For accuracy studies wide confidence intervals for estimates of test accuracy, or true and false positive and negative rates can lower the certainty of evidence.
High probability of Publication bias	Similar criteria to therapy questions A high suspicion of publication bias (e.g., evidence only from small studies supporting a new test, or asymmetry in a funnel plot) can lower the certainty of evidence.
Upgrading for dose effect, large effects residual plausible bias and confounding	Similar criteria to therapy questions For all of these factors, methods have not been fully developed. However, determining a dose effect (e.g., increasing levels of anticoagulation measured by INR increase the likelihood for vitamin K deficiency or vitamin K antagonists). A very high likelihood of disease (not of patient-important outcomes) associated with test results may increase the certainty of the evidence. However, there is some disagreement if and how dose effects play a role in assessing the certainty of evidence in TA studies.

Table 3. Risk of bias criteria of diagnostic accuracy studies derived from QUADAS-2 (48)

Domain	Patient Selection	Index Test	Reference Standard	Flow and Timing
Description	Describe methods of patient selection Describe included patients (previous testing, presentation, intended use of index test, and setting)	Describe the index test and how it was conducted and interpreted	Describe the reference standard and how it was conducted and interpreted	Describe any patients who did not receive the index tests or reference standard or who were excluded from the 2 X 2 table (refer to flow diagram) Describe the interval and any interventions between index tests and the reference standard
Signaling questions (yes, no, or unclear)	Was a consecutive or random sample of patients enrolled? Was a case–control design avoided? Did the study avoid inappropriate exclusions?	Were the index test results interpreted without knowledge of the results of the reference standard? If a threshold was used, was it pre-specified?	Is the reference standard likely to correctly classify the target condition? Were the reference standard results interpreted without knowledge of the results of the index test?	Was there an appropriate interval between index tests and reference standard? Did all patients receive a reference standard? Did all patients receive the same reference standard? Were all patients included in the analysis?
Risk of bias (high, low, or unclear)	Could the selection of patients have introduced bias?	Could the conduct or interpretation of the index test have introduced bias?	Could the reference standard, its conduct, or its interpretation have introduced bias?	Could the patient flow have introduced bias?

Table 4. Indirectness judgment across the body of evidence for true positives.



Outcome: TP

Domain (original question asked	Description	Judgment - Is the evidence sufficiently direct?
Population:	Three studies had high concern because patients were evaluated as inpatients in tertiary care centres; however, we recognize this is how some patients may present in practice.	○ Yes ● Probably yes ○ Probably no ○ No
Intervention: [intervention]	For the index and reference test domains, we considered most studies to have low concern for applicability.	● Yes ○ Probably yes ○ Probably no ○ No
Comparator: [comparison]	For the index and reference test domains, we considered most studies to have low concern for applicability.	● Yes ○ Probably yes ○ Probably no ○ No
Direct comparison	Yes	● Yes ○ Probably yes ○ Probably no ○ No
Outcome: TP	No concerns for test accuracy ratings. Indirectness related to patient outcomes not rated here.	● Yes ○ Probably yes ○ Probably no ○ No
Final judgment about indirectness across domains:	● No indirectness ○ Serious indirectness ○ Very serious indirectness	

Figure 1a and 1b. Basic designs to evaluate tests

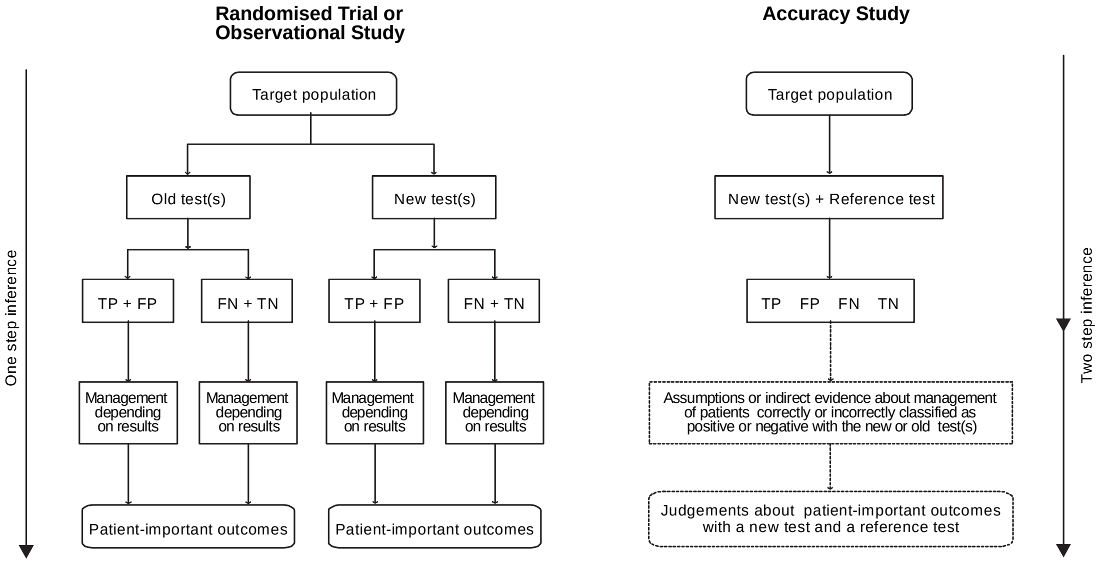

Randomised Trial or Observational Study
Accuracy Study
One step inference
Two step inference
Target population
Old test(s)
New test(s)
TP + FP
FN + TN
TP + FP
FN + TN
Management depending on results
Management depending on results
Management depending on results
Management depending on results
Patient-important outcomes
Patient-important outcomes
Target population
New test(s) + Reference test
TP FP FN TN
Assumptions or indirect evidence about management of patients correctly or incorrectly classified as positive or negative with the new or old test(s)
Judgements about patient-important outcomes with a new test and a reference test

Figure 1 (legend). Two generic ways to evaluate the impact of a test or diagnostic strategy: a) Patients are randomized to either a new test or strategy or to an old test or strategy. Those with a positive test (cases detected) are randomized (or were previously randomized) to receive the best available management (second step of randomization for management not shown in this figure). Investigators evaluate and compare patient-important outcomes in all patients in both groups.(27) b) Patients receive **both** a new test and a reference test (it often, however, is the old or comparator test or strategy). Investigators can then calculate the accuracy of the test compared to the reference test (first step). To make judgments about patient-impact of this test information, patients with a positive test (or strategy) in either group are (or have been in previous studies) submitted to treatment or no treatment; investigators then evaluate and compare patient-important outcomes in all patients in both groups (second step).

Example for Figure 1a - B-type natriuretic peptide for heart failure

Randomized controlled trials (RCTs) that explored a diagnostic strategy guided by the use of B-type natriuretic peptide (BNP) – designed to aid diagnosis of heart failure – compared with no use of BNP in patients presenting to the emergency department with acute dyspnea.(52, 53) As it turned out, the group randomized to receive BNP spent a shorter time in the hospital at lower cost with no increased mortality or morbidity.

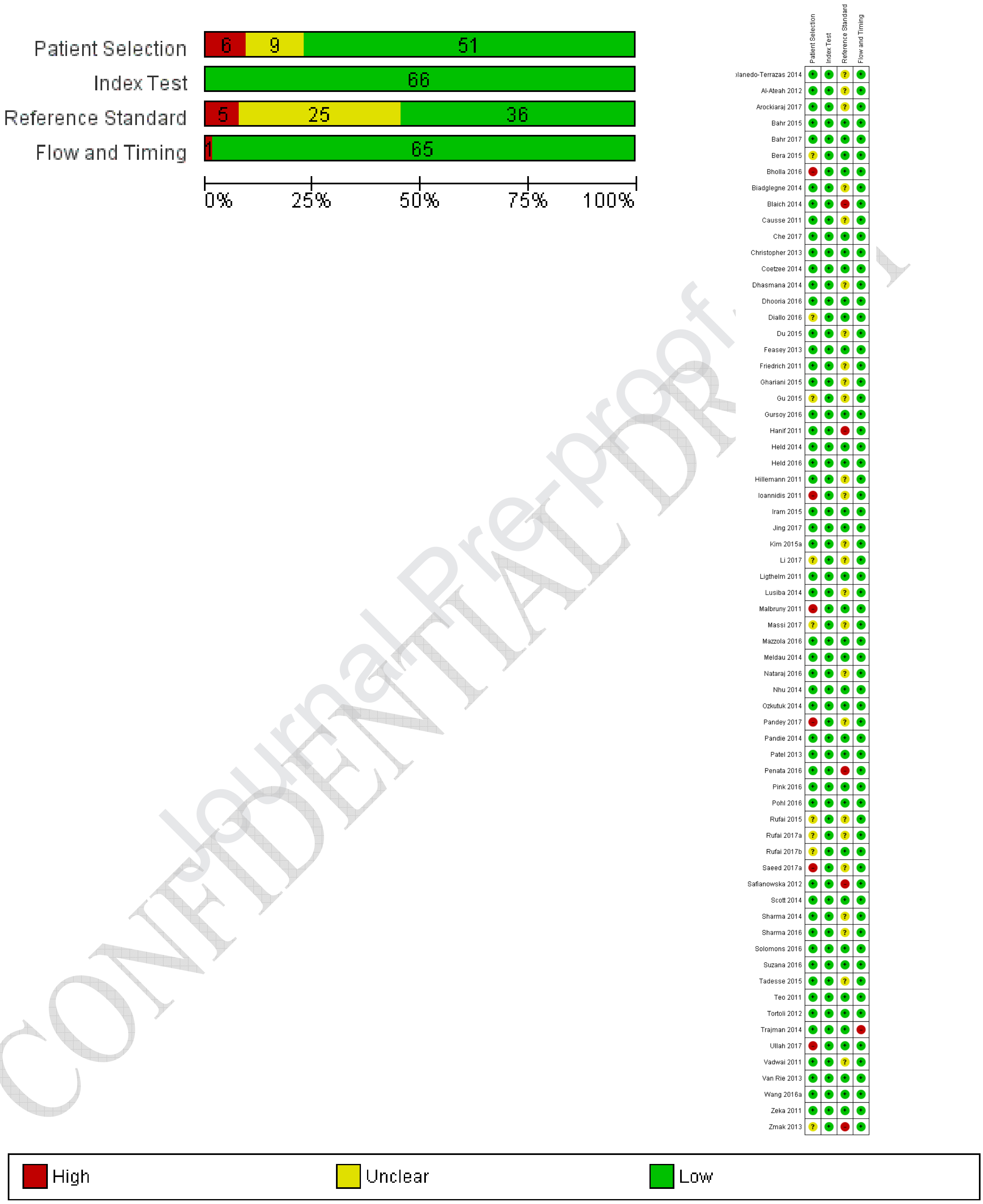
Example for Figure 1b - non-contrast helical CT for urolithiasis

Consistent evidence from well-designed studies demonstrates fewer false negative results with non-contrast helical CT than with intravenous pyelography (IVP) in the diagnosis of suspected acute urolithiasis.(54) However, the stones in the ureter that CT detects but IVP "misses" are smaller, and hence are likely to pass more easily. Since RCTs evaluating the outcomes in patients treated for smaller stones are not available, the extent to which reduction in cases that are missed (false negatives) and follow-up of incidental findings unrelated to renal calculi with CT have important health benefits remains uncertain.(55)

Figure 2. Example of a risk of bias assessment using QUADAS2- from for the use of Xpert in tuberculosis (44)

Figure 2a

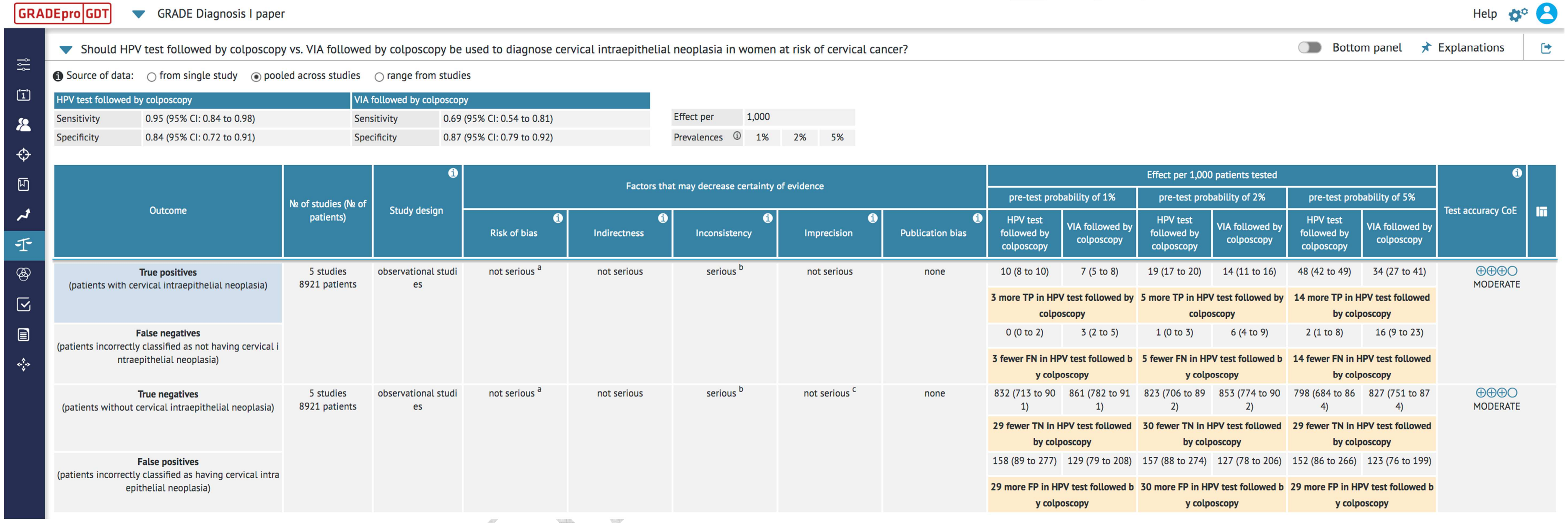
Figure 2b



	Patient Selection	Index Test	Reference Standard	Flow and Timing
lanedo-Terrazas 2014	+	+	?	+
Al-Ateah 2012	+	+	?	+
Arockiaraj 2017	+	+	?	+
Bahr 2015	+	+	+	+
Bahr 2017	+	+	+	+
Bera 2015	?	+	+	+
Bholla 2016	–	+	+	+
Biadglegne 2014	+	+	?	+
Blaich 2014	+	+	–	+
Causse 2011	+	+	?	+
Che 2017	+	+	+	+
Christopher 2013	+	+	+	+
Coetzee 2014	+	+	+	+
Dhasmana 2014	+	+	?	+
Dhooria 2016	+	+	+	+
Diallo 2016	?	+	+	+
Du 2015	+	+	?	+
Feasey 2013	+	+	+	+
Friedrich 2011	+	+	?	+
Ghariani 2015	+	+	?	+
Gu 2015	?	+	?	+
Gursoy 2016	+	+	+	+
Hanif 2011	+	+	–	+
Held 2014	+	+	+	+
Held 2016	+	+	+	+
Hillemann 2011	+	+	?	+
Ioannidis 2011	–	+	?	+
Iram 2015	+	+	+	+
Jing 2017	+	+	+	+
Kim 2015a	+	+	?	+
Li 2017	?	+	?	+
Ligthelm 2011	+	+	+	+
Lusiba 2014	+	+	?	+
Malbruny 2011	–	+	+	+
Massi 2017	?	+	?	+
Mazzola 2016	+	+	+	+
Meldau 2014	+	+	+	+
Nataraj 2016	+	+	?	+
Nhu 2014	+	+	+	+
Ozkutuk 2014	+	+	+	+
Pandey 2017	–	+	?	+
Pandie 2014	+	+	+	+
Patel 2013	+	+	+	+
Penata 2016	+	+	–	+
Pink 2016	+	+	+	+
Pohl 2016	+	+	+	+
Rufai 2015	?	+	?	+
Rufai 2017a	?	+	?	+
Rufai 2017b	?	+	+	+
Saeed 2017a	–	+	?	+
Safianowska 2012	+	+	–	+
Scott 2014	+	+	+	+
Sharma 2014	+	+	?	+
Sharma 2016	+	+	?	+
Solomons 2016	+	+	+	+
Suzana 2016	+	+	+	+
Tadesse 2015	+	+	?	+
Teo 2011	+	+	+	+
Tortoli 2012	+	+	+	+
Trajman 2014	+	+	+	–
Ullah 2017	–	+	+	+
Vadwai 2011	+	+	?	+
Van Rie 2013	+	+	+	+
Wang 2016a	+	+	+	+
Zeka 2011	+	+	+	+
Zmak 2013	?	+	–	+

Figure 2 legend. Figure 2a. Risk of bias graph: review authors' judgements about each domain presented as number of studies and percentages across included studies. Figure 2b. Risk of bias review authors' judgements about each domain for each included study. This risk of bias for the total 66 studies not restricted to tuberculous meningitis included in the review.

Figure 3. Example of expressing comparative accuracy of two tests against a reference test.



Should HPV test followed by colposcopy vs. VIA followed by colposcopy be used to diagnose cervical intraepithelial neoplasia in women at risk of cervical cancer?

Source of data: ○ from single study ● pooled across studies ○ range from studies

HPV test followed by colposcopy		VIA followed by colposcopy	
Sensitivity	0.95 (95% CI: 0.84 to 0.98)	Sensitivity	0.69 (95% CI: 0.54 to 0.81)
Specificity	0.84 (95% CI: 0.72 to 0.91)	Specificity	0.87 (95% CI: 0.79 to 0.92)

Effect per	1,000		
Prevalences	1%	2%	5%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested						Test accuracy CoE
								pre-test probability of 1%		pre-test probability of 2%		pre-test probability of 5%		
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	HPV test followed by colposcopy	VIA followed by colposcopy	HPV test followed by colposcopy	VIA followed by colposcopy	HPV test followed by colposcopy	VIA followed by colposcopy	
True positives (patients with cervical intraepithelial neoplasia)	5 studies 8921 patients	observational studies	not serious [a]	not serious	serious [b]	not serious	none	10 (8 to 10)	7 (5 to 8)	19 (17 to 20)	14 (11 to 16)	48 (42 to 49)	34 (27 to 41)	⊕⊕⊕○ MODERATE
								3 more TP in HPV test followed by colposcopy		**5 more TP in HPV test followed by colposcopy**		**14 more TP in HPV test followed by colposcopy**		
False negatives (patients incorrectly classified as not having cervical intraepithelial neoplasia)								0 (0 to 2)	3 (2 to 5)	1 (0 to 3)	6 (4 to 9)	2 (1 to 8)	16 (9 to 23)	
								3 fewer FN in HPV test followed by colposcopy		**5 fewer FN in HPV test followed by colposcopy**		**14 fewer FN in HPV test followed by colposcopy**		
True negatives (patients without cervical intraepithelial neoplasia)	5 studies 8921 patients	observational studies	not serious [a]	not serious	serious [b]	not serious [c]	none	832 (713 to 901)	861 (782 to 911)	823 (706 to 892)	853 (774 to 902)	798 (684 to 864)	827 (751 to 874)	⊕⊕⊕○ MODERATE
								29 fewer TN in HPV test followed by colposcopy		**30 fewer TN in HPV test followed by colposcopy**		**29 fewer TN in HPV test followed by colposcopy**		
False positives (patients incorrectly classified as having cervical intraepithelial neoplasia)								158 (89 to 277)	129 (79 to 208)	157 (88 to 274)	127 (78 to 206)	152 (86 to 266)	123 (76 to 199)	
								29 more FP in HPV test followed by colposcopy		**30 more FP in HPV test followed by colposcopy**		**29 more FP in HPV test followed by colposcopy**		

Outcome: TP

Domain (original question asked	Description	Judgment - Is the evidence sufficiently direct?
Population:	Three studies had high concern because patients were evaluated as inpatients in tertiary care centres; however, we recognize this is how some patients may present in practice.	□ Yes ☒ Probably yes □ Probably no □ No
Intervention: [intervention]	For the index and reference test domains, we considered most studies to have low concern for applicability.	☒ Yes □ Probably yes □ Probably no □ No
Comparator: [comparison]	For the index and reference test domains, we considered most studies to have low concern for applicability.	☒ Yes □ Probably yes □ Probably no □ No
Direct comparison	Yes.	☒ Yes □ Probably yes □ Probably no □ No
Outcome: TP	No concerns for test accuracy ratings, indirectness related to patient outcomes is not rated here in the assessment of certainty that focuses on accuracy alone.	☒ Yes □ Probably yes □ Probably no □ No
Final judgment about indirectness across domains:	☒ No indirectness □ Serious indirectness □ Very serious indirectness	

Key findings

Rating the certainty of the body of evidence (quality of evidence or confidence in estimates) of test accuracy studies differs conceptually but shares the fundamental logic for the domains risk of bias and indirectness of the GRADE approach for intervention, prognostic or other studies.

What this adds to what is known?

Questions about the relative merit of alternative testing strategies in clinical and public health require framing in terms of health outcomes. Evidence evaluation will often, however, begin with an evidence synthesis - ideally a systematic review or health technology assessment - and rating of test accuracy, and subsequently move to evaluation of the evidence linking test accuracy to patient-important and population outcomes. We describe examples for how GRADE has been applied to test accuracy studies in Cochrane and other reviews and World Health Organization and other guidelines focusing on risk of bias and indirectness in part 1 of this article.

What are the implications, what should change now?

Investigators interested in using the GRADE for diagnostic and other healthcare related tests should consider the guidance offered in this article about how to evaluate research that focuses on the impact of tests, specifically from test accuracy studies in the context of risk of bias and indirectness. We provide examples for how to separate indirectness on a population, test intervention, test comparison and outcome levels.

Disclosure Statement

The authors are members of the GRADE Working Group. They have made various contributions to the development of its methods. HR reports: As part of my employment with Kleijnen Systematic Reviews Ltd. I have been working on projects for Bayer and Grunenthal.

CREDIT statement
Holger J Schünemann: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Software; Supervision; Visualization; Roles/Writing – original draft; Writing – review & editing.

Reem A. Mustafa: Conceptualization; Investigation; Methodology; Writing – review & editing.

Jan Brozek: Conceptualization; Investigation; Methodology; Writing – review & editing.

Karen R Steingart: Investigation; Methodology; Visualization; Writing – review & editing.

Mariska Leeflang: Conceptualization; Investigation; Methodology; Writing – review & editing.

Mohammad Hassan Murad: Conceptualization; Investigation; Methodology; Visualization; Writing – review & editing.

Patrick Bossuyt: Conceptualization; Investigation; Methodology; Writing – review & editing.

Paul Glasziou: Conceptualization; Investigation; Methodology; Writing – review & editing.

Roman Jaeschke: Methodology; Writing – review & editing.

Stefan Lange: Conceptualization; Investigation; Methodology; Writing – review & editing.

Joerg Meerpohl: Conceptualization; Investigation; Methodology; Writing – review & editing.

Miranda Langendam: Investigation; Methodology; Writing – review & editing.

Monica Hultcrantz: Investigation; Methodology; Writing – review & editing.

Gunn E Vist: Conceptualization; Investigation; Methodology; Writing – review & editing.

Elie A Akl: Conceptualization; Investigation; Methodology; Writing – review & editing.

Mark Helfand: Investigation; Methodology; Writing – review & editing.

Nancy Santesso: Conceptualization; Investigation; Methodology; Writing – review & editing.

Lotty Hooft: Investigation; Methodology; Writing – review & editing.

Rob Scholten: Investigation; Methodology; Writing – review & editing.

Måns Rosen: Investigation; Methodology; Writing – review & editing.

Anne Rutjes: Investigation; Methodology; Writing – review & editing.

Mark Crowther: Investigation; Methodology; Writing – review & editing.

Paola Muti: Conceptualization; Investigation; Writing – review & editing.

Heike Raatz: Conceptualization; Investigation; Writing – review & editing.

Mohammed T. Ansari: Conceptualization; Methodology; Investigation; Writing – review & editing.

John Williams: Conceptualization; Investigation; Methodology; Writing – review & editing.

Regina Kunz: Conceptualization; Investigation; Methodology; Writing – review & editing.

Jeff Harris: Conceptualization; Investigation; Writing – review & editing.

Ingrid Arévalo Rodriguez: Investigation; Writing – review & editing.

Mikashmi Kohli; Investigation; Methodology; Visualization; Writing – review & editing.

Gordon H Guyatt; Conceptualization; Investigation; Methodology; Writing – review & editing.