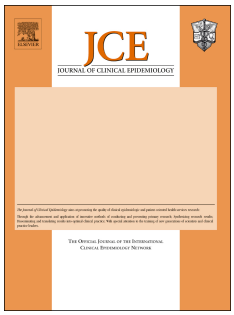


GRADE guidelines: 21 part 2. Inconsistency, Imprecision, publication bias and other domains for rating the certainty of evidence for test accuracy and presenting it in evidence profiles and summary of findings tables

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)30674-2

DOI: https://doi.org/10.1016/j.jclinepi.2019.12.021

Reference: JCE 10047

To appear in: *Journal of Clinical Epidemiology*

Received Date: 27 July 2019

Revised Date: 28 November 2019

Accepted Date: 30 December 2019

Please cite this article as: Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, Bossuyt P, Glasziou P, Jaeschke R, Lange S, Meerpohl J, Langendam M, Hultcrantz M, Vist GE, Akl EA, Helfand M, Santesso N, Hooft L, Scholten R, Rosen M, Rutjes A, Crowther M, Muti P, Raatz H, Ansari MT, Williams J, Kunz R, Harris J, Rodriguez IA, Kohli M, Guyatt GH, for the GRADE Working Group, 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, *Journal of Clinical Epidemiology* (2020), doi: https://doi.org/10.1016/j.jclinepi.2019.12.021.

during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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

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: 2826

Tables: 2

Figures: 5

Key findings

Rating the certainty of the body of evidence (quality of evidence or confidence in estimates) on the domains imprecision, inconsistency and publication bias for test accuracy studies shares the fundamental logic of the GRADE approach for intervention, prognostic or other studies but requires different operationalization.

What this adds to what is known?

Evidence evaluation will often begin with an evidence synthesis - ideally a systematic review or health technology assessment - and the rating of certainty in test accuracy includes assessing inconsistency, imprecision, publication bias and other domains. In this part 2 of GRADE guidance 21, we describe the judgments on these domains and across a body of evidence using examples from how GRADE has been applied to test accuracy studies in Cochrane and other reviews as well as World Health Organization and other guidelines.

What are the implications, what should change now?

Further work is needed for better operationalization of the domain imprecision and domains that may lead to increasing the certainty. However, investigators interested in using the GRADE for diagnostic and healthcare related tests should consider the guidance offered in this article for the corresponding domains and how the information is presented in evidence profiles and summary of findings tables.

Abstract

Objectives: This article provides updated GRADE guidance about how authors of systematic reviews and health technology assessments (HTA) and guideline developers can rate 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) on the domains imprecision, inconsistency, publication bias and other domains. It also provides guidance for how to present synthesized information in evidence profiles and summary of findings tables.

Study Design and Setting: We present guidance for rating certainty in TA in clinical and public health and review the presentation of results of a body of evidence regarding tests.

Results: Supplemented by practical examples, we describe how raters of the evidence can apply the GRADE domains inconsistency, imprecision, and publication bias to a body of evidence of TA studies.

Conclusions: Using GRADE in Cochrane and other reviews as well as World Health Organization and other guidelines helped refining the GRADE approach for rating the certainty of a body of evidence from TA studies. While several of the GRADE domains (e.g., imprecision and magnitude of the association) require further methodological research to help operationalize them, judgments need to be made on the basis of what is known so far.

Key words: GRADE, diagnosis, tests, test accuracy, certainty of evidence, diagnostic accuracy, guidelines, systematic reviews, HTA

GRADE guidelines: 21 part 2. Inconsistency, imprecision, publication bias and other domains for rating the certainty and presenting evidence profiles and summary of findings tables

1.0 Introduction

In part 1 of this 21st article in the GRADE guidance series in the Journal of Clinical Epidemiology we described the unique challenges about rating the initial study designs, risk of bias and indirectness in studies about test accuracy (TA).(1) We also introduced three examples of questions about the use of tests to which we will refer to also in this article (Box 1).(1-3) In this part 2 of GRADE guidance 21, we will 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 focusing on the domains inconsistency, imprecision, publication bias and domains that may increase our certainty. With regards to HTA, we refer to the rating of the certainty of TA results from a body of evidene that may be used for other aspects of HTA, such as modelling and cost analyses. This article also describes how authors of SR and HTA can present the results of an assessment to decision makers and it supplements our previous work addressing GRADE for diagnostic questions and the GRADE Evidence to Decision Frameworks for tests.(2, 4, 5)

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?(6)***

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.(7)***

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

2.0 The GRADE certainty domains inconsistency, imprecision and publication bias

We continue our description of the rating of the certainty by domain and across domains from part 1 of this article beginning with inconsistency.(1)

2.1. Certainty of the evidence - inconsistency

Important unexplained inconsistency of the results across studies may decrease certainty of the evidence. Raters should evaluate estimates for sensitivity and specificity separately. As for intervention studies, raters should consider meta-analyses when the evidence would support them. Judgments of the extent of heterogeneity are based on similarity of the point estimates, extent of overlap of confidence intervals, variance estimates in random effects meta-analysis and statistical criteria including tests of heterogeneity. Any role of I^2 in assessing heterogeneity in meta-analyses of TA requires further exploration in TA studies but it comes with similar limitation as in intervention studies.

2.1.1. *Examples for inconsistency*

In our tuberculous meningitis example, although the sensitivity estimates ranged from 33% to 100%, the absence of concentrating the sample in preparing the CSF specimen in certain settings could explain some of the heterogeneity (higher sensitivity in concentrated samples).(8) Overall the confidence intervals were overlapping for all but one study (Figure 1). Specificity was similar across the studies. The raters did not lower the certainty for inconsistency (Table 1).

Insert Table 1 approximately here

Steingart and colleagues' conducted a systematic review evaluating commercial serological tests for the diagnosis of pulmonary and extrapulmonary tuberculosis. For extrapulmonary tuberculosis, they reported sensitivity values from 0% to 100%, and specificity values from 59% to 100%. This variability was sufficiently great that, in the presence of non-overlapping confidence intervals and limited explanation for the inconsistency (e.g. identity of the commercial test, antibody detected and site of extrapulmonary TB), the authors chose not to derive summary accuracy estimates (Figure 2) and rated down for inconsistency.(9) While the previous two examples represent debatable exercises of judgment, less challenging examples exist. For instance, consider the investigation of the accuracy of T-SPOT.TB, an interferon-gamma release assay, for active tuberculosis in people presumed to have tuberculosis without HIV infection (Figure 3). Here, similar point estimates and overlapping confidence intervals support the judgment not to rate down for inconsistency.

When differences in the populations enrolled, the index test or reference test applied, or the outcomes measured, explain inconsistency in the test accuracy estimates, presenting results in subgroups is often appropriate. Variability in the investigators' choice of test

thresholds may, for instance, explain heterogeneity and be elucidated in an receiving operarator characteristics (ROC) analysis. Ideally, inconsistency should be assessed by using clearly defined thresholds that either resemble healthcare practice or will be used to guide practice. For example, the variability in thresholds used to describe pleural effusion as being of cardiac origin based on pro-brain natriuretic peptide (pro-BNP), explained some of the inconsistency in the sensitivity and specificity observed in a systematic review evaluating the utility of this test.(10)

2.2. Certainty of the evidence - imprecision

Wide confidence intervals for estimates of test accuracy or true and false positive and negative rates can reduce the certainty of the evidence or diagnostic odds ratio (DOR) or another accuracy measure. Here, we focus on the confidence intervals (CI) around sensitivity and specificity (note that the imprecision may also be expressed as credible intervals). What is wide enough to rate down is, however, a matter of judgement, and these decisions may vary depending on the context.(11)

For systematic review authors, imprecision judgments can be based on both the width of the confidence interval (CI) and the number of participants in the studies. The CI depend on the number of events; for sensitivity it is the number of diseased persons and the number of test positives; for the specificity it is the number of non-diseased persons and the number of test-negatives. In contextualized settings, i.e. when decision making is influenced by weighing the TP, FN, TN and FP against each other and the downstream consequences, raters should set thresholds for the confidence intervals that reflect the implications for people or patient management. When the boundaries of the CI include values that may lead

to different conclusions of the test's value the certainty of the evidence may be lowered. This implies that a relatively narrow CI may still be too wide to make a firm conclusion. For example, if review authors or a guideline panel agrees beforehand that a sensitivity of 0.8 would be the lowest acceptable sensitivity for a certain situation, then a CI that runs from 0.72 to 0.88 may be too wide to conclude whether use of the test provides more benefits than harms. On the other hand, a CI between 0.82 and 0.92 may be considered narrow enough to draw a conclusion. For decision makers, this should be done by translating the estimates for sensitivity and specificity (and their confidence intervals) to absolute numbers of TP, FP, FN, TN (and any upper and lower limits around these) for assumed prevalences. For example, if the average number of people tested per year for a condition is 1000, and the expected prevalence among this population is to be 1%, then 10 people are expected to have the disease and in that case a wider CI around sensitivity may lead to less concern about imprecision than when the prevalence is around 40% because a fairly narrow confidence interval for sensitivity may lead to a wide CI for the TP.

2.2.1. Examples for imprecision

Based on the pooled estimates of sensitivity and specificity, Kohli and colleagues calculated the estimates of TP, FN, TN and FP for different prevalences of tuberculous meningitis (Table 1). Based on 433 patients in 29 studies, they judged the limits of the credible intervals for the TP and FN to be sufficiently wide to warrant rating down for imprecision; in contrast, they found that the limits for the TN and FP were sufficiently narrow that rating down was unnecessary. For our example about diagnosis of CMA (Table 2), the CI were sufficiently narrow not to warrant downgrading.

2.3. Certainty of the evidence - Publication bias

Generally, raters should make publication bias judgments using the same criteria as in therapeutic studies: for-profit interest, the presence of only studies that produce precise estimates of high accuracy despite small sample size, and knowledge about studies that were conducted but are not published. Although high suspicion of publication bias will decrease our certainty in TA results, little is known about the actual existence of publication bias in TA studies. Applying widely used tests for funnel plot asymmetry (e.g Egger's or Begg's tests) in test accuracy systematic reviews is likely to result in rating down for publication bias more frequently than appropriate. For instance, study size may correlate with test accuracy as a result of patients' or study characteristics rather than publication bias.

Other tests (e.g. Deeks' test or the trim and fill method) may be more appropriate for testing publication bias in TA systematic reviews. (12-14) The trim and fill method, in particular, has advantages that include providing an estimate of the unbiased TA and an intuitive visual display that includes both the observed studies and the imputed studies, allowing authors to visually inspect how much TA changes when the imputed studies are included. If this change is trivial, then there is no need to rate down certainty for publication bias.

A special situation of publication bias may occur with non-inferiority test accuracy studies. In that case, accuracy of a new index test compared with the reference test is based on the difference in the paired partial area under the ROC curve. One can test this difference with Bayesian statistical methods that result in assessing statistical significance.(15) Because of

the ability to assess statistical significance, this design may be more susceptible to not publishing negative findings, which theoretically can lead to publication bias. Given the limitations of all the available statistical models and methods to test for publication bias in TA studies, confident inferences that publication bias exists may be restricted to knowledge that unpublished TA studies exist. The lack of a standardized method to register TA studies, however, makes such knowledge difficult to obtain.

2.3.1 Examples for publication bias

Kohli and coauthors did not rate down for publication bias despite concerns about for-profit interest and small studies. This was due to the comprehensiveness of the literature search and the extensive outreach to TB researchers did not identify unpublished studies.

In a systematic review that assessed the accuracy of magnetic resonance imaging (MRI) in identifying liver iron overload in patients with hereditary hemochromatosis, hemoglobinopathy, or myelodysplastic syndrome the authors suspected publication bias. (16) In the regression test for funnel plot asymmetry (Deek's test), the P-value for the slope coefficient was 0.07. (16) Murad and colleagues also used Deeks' funnel plot asymmetry tests and visual inspection of funnel plots in their review of Fractional Exhaled Nitric Oxide (FeNO) in Asthma Management. The authors described potential publication bias for cutoffs<20, and no indication of publication bias for cutoffs 20-30 (Figure 4).(17)

3.0. Certainty of the evidence - upgrading for test-outcome relations, large estimates of TA and residual plausible bias and confounding

Upgrading may be relevant for rating a body of evidence from studies of TA. Certainty in TA may increase if the ROC curve shows a clear and consistent sensitivity-specificity

relationship (the diagnostic equivalent of a dose effect). The strongly increased likelihood of acute myocardial infarction with increasing levels of troponin T increases our confidence in the diagnostic properties of this test, that is a strong correlation between increasing test values and the likelihood and severity of disease as opposed to the mathematical phenomenon of a simple increase in the likelihood because of choosing different cut-offs for the test values.(18) Very high accuracy of a test, and the presence of minimal opposing residual confounding (19) might also increase one's confidence in the usefulness of the test. Compared with the effects on the certainty in therapeutic interventions in observational studies, the methods to determine whether the evidence warrants rating up on a particular domain is, however, less well established for tests and requires further theoretical and empirical work. Even amongst the authors of this article, there is no agreement if and how, for example, dose-effects play a role in assessing the certainty in estimates in TA studies.

3.1. Examples for upgrading

For example, evidence suggesting a threshold dependent identification of false negatives and false positives in the diagnosis of asthma with FeNO may increase the certainty of the test accuracy studies (Figure 5).

4.0. Arriving at a bottom line for the certainty of the evidence

Tables 1 and 2 show the assessment of the certainty in the evidence of TA and the summary of findings (SoF) table for examples 2 and 3 (Box 1). Kohli and colleagues rated down the certainty for TP and FN (sensitivity) for imprecision but not for TN and FP (specificity) (Table 1). The included accuracy studies were well planned and executed, the systematic review

authors undertook investigations to explain inconsistency, and there was little reason for a high suspicion of publication bias. The authors judged, however, that credible intervals for the sensitivity of the test were excessively wide and rated the overall certainty moderate for sensitivity and high for specificity. This example demonstrates that the certainty frequently differs for the accuracy outcome pairs TP and FN (sensitivity) compared to the FP and TN (specificity).

In the example evaluating tests for CMA (Table 2), most studies enrolled highly selected patients with atopic eczema or gastrointestinal symptoms, no study reported if an index test or a reference standard were interpreted without knowledge of the results of the other test and it is very likely that those interpreting results of one test knew the results of the other. In addition, all except for one study that reported withdrawals did not explain why patients were withdrawn. The systematic review authors, therefore, rated down for risk of bias. They also rated down for imprecision for TP, FN, TN and FP for an overall rating of low certainty.

5.0 Presentation of results – Evidence Profiles and Summary of Findings Tables

Clear presentations of information about diagnostic tests in evidence summaries helps ensure transparency for decision-makers. We described evidence profiles and SoF tables in prior articles in this series.(20) When the focus in on TA, the presentation format differs from presentation of questions about therapy or interventions.

GRADE identified three types of layers of evidence summaries that might be useful and can be developed in GRADE's official app GRADEpro: First, simple SoF tables and evidence profiles that provide information about TA alone (we refer to this as layer 1, illustrated in

Tables 1 and 2). Tables may also include basic information regarding other features related to the test or test strategy facilitating decision-making such as direct complications of a test that can be derived from accuracy studies (layer 2). Tables that provide the information as patient-important outcomes and include explicit judgments about the desirable and undesirable health effects of tests (layer 3) are useful during the process of making recommendations.(3) The format of Tables 1 and 2 and the corresponding interactive SoFs (iSoFs) are based on the results of testing of alternative presentation formats with various user groups.(21) Systematic review authors may sometimes use layer 2 to describe the direct consequences of a test apart from TA. For example, the direct undesirable effects of a test such as anaphylactic reactions from radiological contrast dye, inconclusive results or direct burden from the test may be described to facilitate decision making.

Layer 3 includes information for health outcomes following a decision analysis of the various scenarios that result in patient or population important outcomes. We described layer 3 in other articles in this series that address recommendations about diagnostic and other tests and strategies.(2, 3)

Limitations of Level 1 SoF Table format includes challenges in managing continuous or multi-level tests easily. These results are best presented through interactive SoF (iSoF) tables in GRADEpro to which we provide hyperlinks in tables 1 and 2. The separate columns for true/false positive/negatives are most useful for an analysis of consequences of test outcomes (when indirectness is a problem) but also introduce redundancy. However, our user testing suggests that the current format is helpful in summarizing results in systematic reviews and included presentation in the setting of guideline panels.

6.0 Conclusion

The GRADE approach to rating the certainty of evidence for TA is comprehensive and transparent. We have presented an overview of this approach, provided examples and reviewed the presentation of results of a body of evidence for TA studies in this and the prior article (part 1). Although several of the domains (e.g., imprecision, publication bias and magnitude of the association) will benefit from further elaboration in methodological research, they have been applied in many systematic reviews and guidelines.

In the next article 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.(3) We will also provide alternative ways of presenting this information during the development of recommendations and to users of guidelines. In addition, in another article, the GRADE working group provided a conceptual approach to defining the certainty of evidence for test accuracy studies.(22)

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.

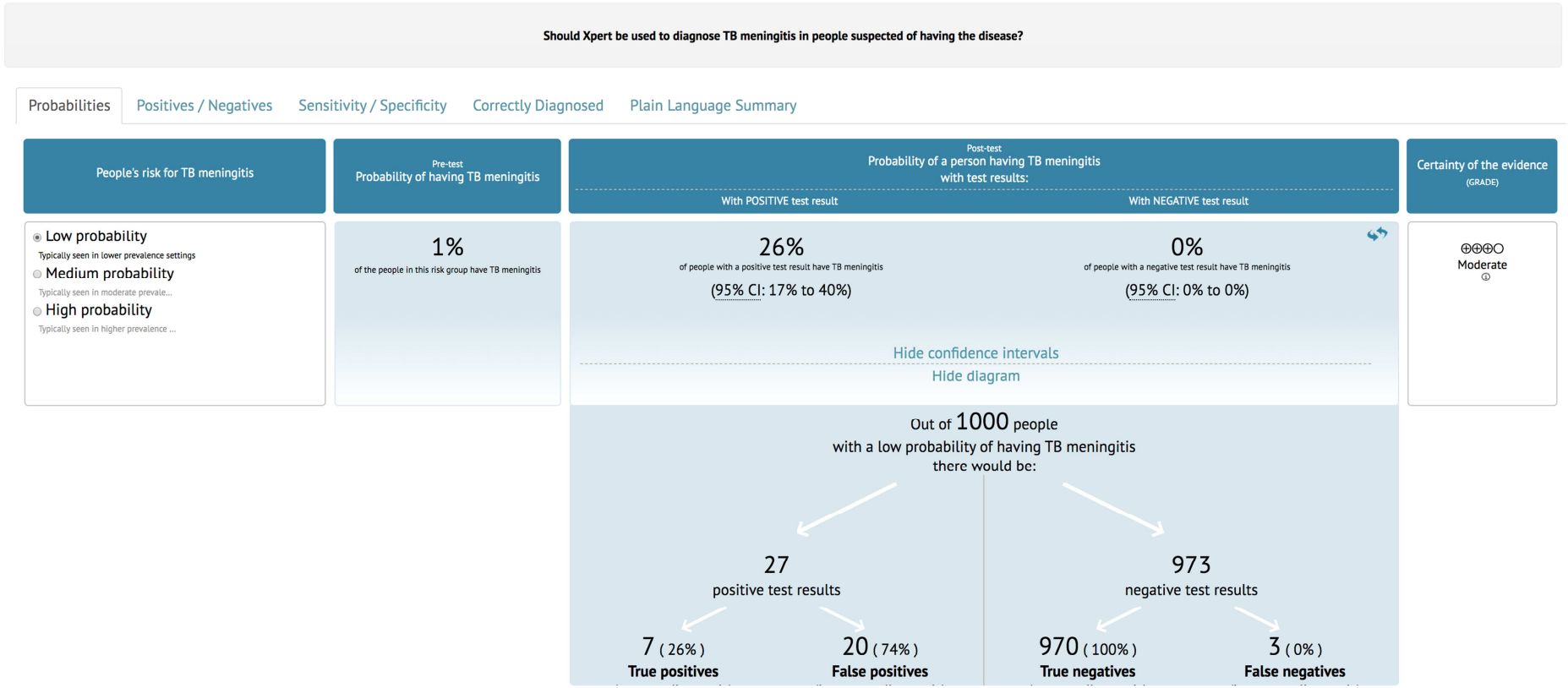
References

1. Schunemann HJ, Mustafa RA, Brozek J, Steingart K, Leeflang M, Murad HM, et al. GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. in press.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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).
10. Janda S, and Swiston, J. Diagnostic accuracy of pleural fluid NT-pro-BNP for pleural effusions of cardiac origin: A systematic review and meta-analysis. BMC Pulmonary Medicine. 2010;10(58).
11. Schunemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol. 2016;75:6-15.
12. 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(9):882-93.
13. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455-63.
14. Burkner PC, Doebler P. Testing for publication bias in diagnostic meta-analysis: a simulation study. Stat Med. 2014;33(18):3061-77.
15. Li CR, Liao CT, Liu JP. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. Stat Med. 2008;27(10):1762-76.

16. Sarigianni M, Liakos A, Vlachaki E, Paschos P, Athanasiadou E, Montori VM, et al. Accuracy of magnetic resonance imaging in diagnosis of liver iron overload: a systematic review and meta-analysis. Clin Gastroenterol Hepatol. 2015;13(1):55-63 e5.
17. Wang Z, Pianosi PT, Keogh KA, Zaiem F, Alsawas M, Alahdab F, et al. The Diagnostic Accuracy of Fractional Exhaled Nitric Oxide Testing in Asthma: A Systematic Review and Meta-analyses. Mayo Clin Proc. 2018;93(2):191-8.
18. Hill SA, Devereaux PJ, Griffith L, Opie J, McQueen MJ, Panju A, et al. Can troponin I measurement predict short-term serious cardiac outcomes in patients presenting to the emergency department with possible acute coronary syndrome? CJEM. 2004;6(1):22-30.
19. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6.
20. 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.
21. Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, et al. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PLoS One. 2015;10(10):e0134553.
22. Hultcrantz M, Mustafa RA, Leeflang MMG, Lavergne V, Estrada-Orozco K, Ansari MT, et al. Defining ranges for certainty ratings of diagnostic accuracy: A GRADE concept paper. J Clin Epidemiol. 2019.
23. 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. Implementation Science. 2011;6:62.
24. Metcalfe JZ, Everett CK, Steingart KR, Cattamanchi A, Huang L, Hopewell PC, et al. Interferon-gamma release assays for active pulmonary tuberculosis diagnosis in adults in low- and middle-income countries: systematic review and meta-analysis. J Infect Dis. 2011;204 Suppl 4:S1120-9.

Table 1. Summary of findings table and evidence profile summarizing diagnostic test accuracy studies informing the question “Should Xpert be used to diagnose TB meningitis in people suspected of having the disease? (8)” Non-contextualized certainty of the evidence in test accuracy rating (without rating the indirectness stemming from the link between accuracy data and patient outcomes). For an interactive version in GRADEpro see this hyperlink iSoF Table 1 (also including a plain language summary)

Summary of Findings Table: Should Xpert be used to diagnose TB meningitis in people suspected of having the disease?



Evidence profile: Should Xpert be used to diagnose TB meningitis in people suspected of having the disease?

Sensitivity	0.71 (95% CI: 0.61 to 0.80)
Specificity	0.98 (95% CI: 0.97 to 0.99)

Prevalences	1%	5%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of 1%	pre-test probability of 5%	pre-test probability of 10%	
True positives (patients with TB meningitis)	29 studies 433 patients	cross-sectional (cohort type accuracy study)	not serious [a]	not serious [b]	not serious [c]	serious [d]	none	7 (6 to 8)	36 (30 to 40)	71 (61 to 80)	⊕⊕⊕◯ MODERATE
False negatives (patients incorrectly classified as not having TB meningitis)								3 (2 to 4)	14 (10 to 20)	29 (20 to 39)	
True negatives (patients without TB meningitis)	29 studies 3341 patients	cross-sectional (cohort type accuracy study)	not serious	not serious	not serious	not serious	none	970 (960 to 978)	931 (922 to 939)	882 (873 to 889)	⊕⊕⊕⊕ HIGH
False positives (patients incorrectly classified as having TB meningitis)								20 (12 to 30)	19 (11 to 28)	18 (11 to 27)	

Explanations

a. As assessed by QUADAS-2, for the reference standard domain there were only four studies (14%) that had unclear risk of bias because specimens underwent decontamination. We did not downgrade.

b. For indirectness, regarding applicability, for the patient selection domain, we considered most studies to have unclear concern. 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. For the index and reference test domains, we considered most studies to have low concern for applicability. We did not downgrade.

c. For individual studies, sensitivity estimates ranged from 33% to 100%. We thought that low TB prevalence and absence of concentration in preparing the cerebrospinal fluid specimen could explain some of the heterogeneity in sensitivity results. We did not downgrade.

d. The wide CrI around true positives and false negatives may lead to different decisions depending on which credible limits are assumed. We downgraded one level.

Table 2. Summary of Findings table and evidence profile summarizing diagnostic test accuracy studies informing the question "Should skin prick tests be used for the diagnosis of IgE-mediated cow's milk allergy (CMA) in patients suspected of CMA?". Non-contextualized certainty of the evidence in test accuracy rating (without rating the indirectness stemming from the link between accuracy data and patient outcomes).(23) For an interactive version in GRADEpro with case descriptors see this hyperlink iSoF Table 2 (also including a plain language summary)

Summary of Findings Table: Should skin prick tests be used to diagnose IgE-mediated cow's milk allergy (CMA) in patients suspected of CMA?
Patient or population : patients suspected of CMA
Setting : children suspected of IgE-mediated CMA
New test : [comparator test] | **Cut-off value** :
Reference test : oral food challenge | **Threshold** : anaphylaxis, burden on time and anxiety for family, exclusion of milk and use of special formula
Pooled sensitivity : 0.67 (95% CI: 0.64 to 0.70) | **Pooled specificity** : 0.74 (95% CI: 0.72 to 0.77)

Test result	Number of results per 1,000 patients tested (95% CI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 10% Typically seen in	**Prevalence 40%** Typically seen in	**Prevalence 80%** Typically seen in		
True positives	**67** (64 to 70)	**268** (256 to 280)	**536** (512 to 560)	2302 (23)	⊕⊕◯◯ **LOW** [a,b]
False negatives	**33** (30 to 36)	**132** (120 to 144)	**264** (240 to 288)		
True negatives	**666** (648 to 693)	**444** (432 to 462)	**148** (144 to 154)	2302 (23)	⊕⊕◯◯ **LOW** [a,b]
False positives	**234** (207 to 252)	**156** (138 to 168)	**52** (46 to 56)		

CI: Confidence interval

Evidence Profile

Question: Should skin prick tests be used to diagnose IgE-mediated cow's milk allergy (CMA) in patients suspected of CMA?

Sensitivity	0.67 (95% CI: 0.64 to 0.70)
Specificity	0.74 (95% CI: 0.72 to 0.77)

Prevalences	10%	40%	80%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of 10%	pre-test probability of 40%	pre-test probability of 80%	
True positives (patients with IgE-mediated	23 studies 2302	cross-sectional (cohort type	serious [a]	not serious	serious [b]	not serious	none	67 (64 to 70)	268 (256 to 280)	536 (512 to 560)	⊕⊕◯◯

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of 10%	pre-test probability of 40%	pre-test probability of 80%	
cow's milk allergy (CMA))	patients	accuracy study)									LOW
False negatives (patients incorrectly classified as not having IgE-mediated cow's milk allergy (CMA))								33 (30 to 36)	132 (120 to 144)	264 (240 to 288)	
True negatives (patients without IgE-mediated cow's milk allergy (CMA))	23 studies 2302 patients	cross-sectional (cohort type accuracy study)	serious [a]	not serious	serious [b]	not serious	none	666 (648 to 693)	444 (432 to 462)	148 (144 to 154)	⊕⊕◯◯ LOW
False positives (patients incorrectly classified as having IgE-mediated cow's milk allergy (CMA))								234 (207 to 252)	156 (138 to 168)	52 (46 to 56)	

Explanations

a. Most studies enrolled highly selected patients with atopic eczema or gastrointestinal symptoms, no study reported if an index test or a reference standard were interpreted without knowledge of the results of the other test, but it is very likely that those interpreting results of one test knew the results of the other; all except for one study that reported withdrawals did not explain why patients were withdrawn.

b. Estimates of sensitivity ranged from 10% to 100%, and specificity from 14% to 100%; we could not explain it by quality of the studies, tests used or included population.

Figure 1. Forest plots of Xpert sensitivity and specificity for tuberculous meningitis.(8)

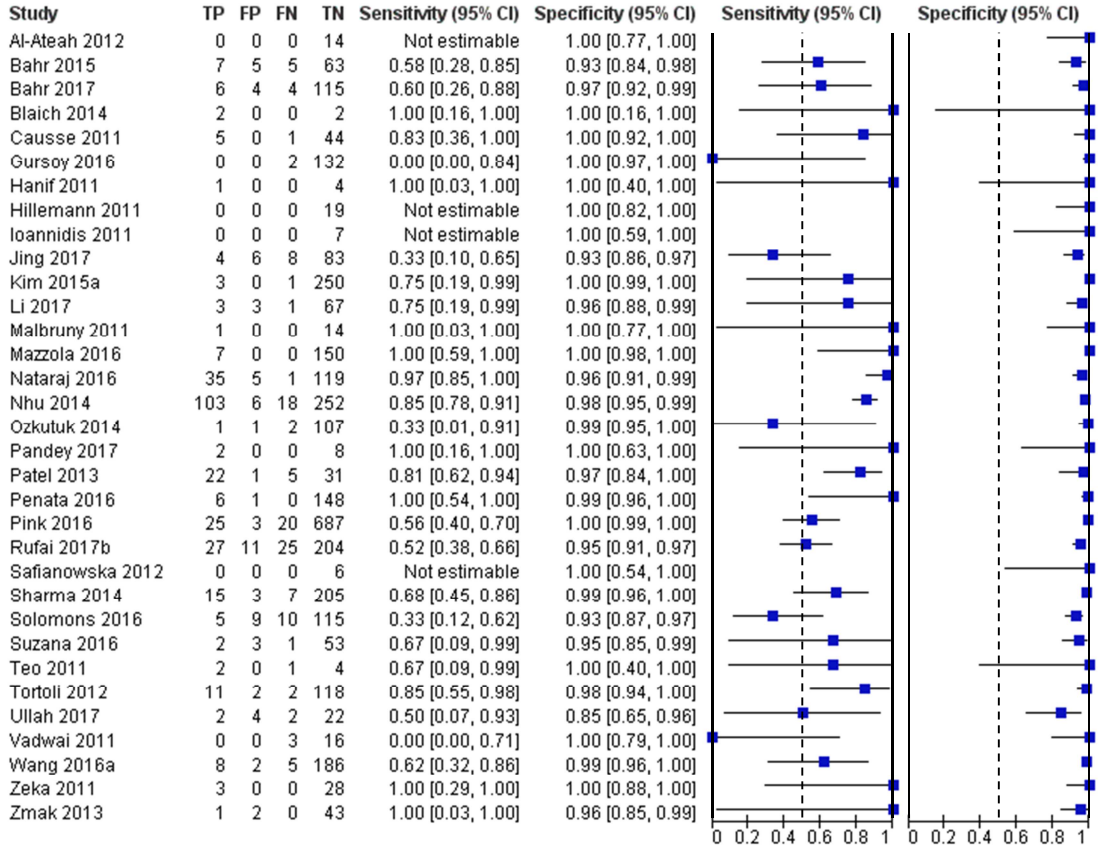
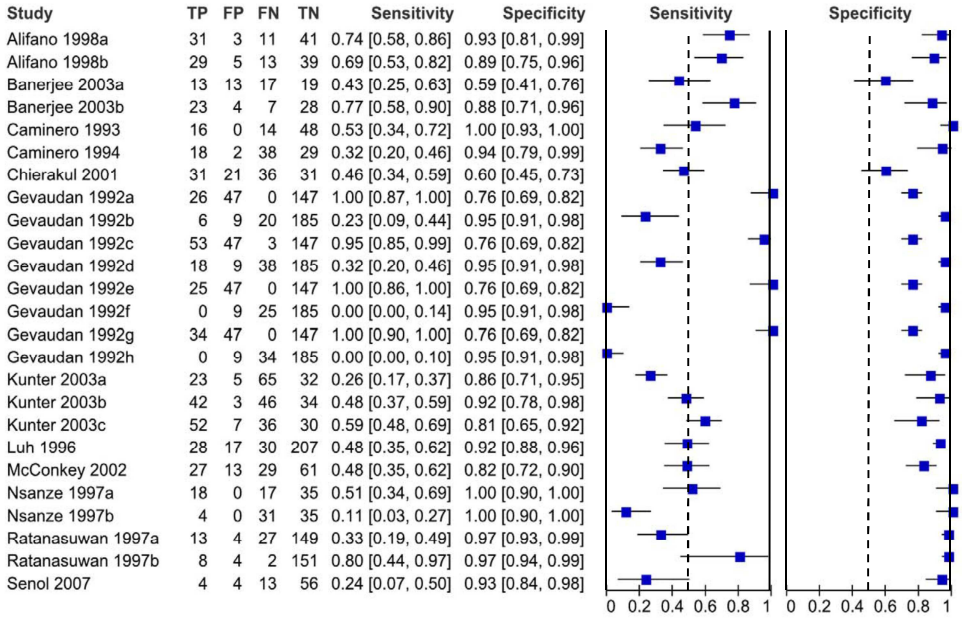


Figure 2. Forest plots of sensitivity and specificity of commercial serological tests for extrapulmonary TB, all studies.(9)



Study	TP	FP	FN	TN	Sensitivity	Specificity
Alifano 1998a	31	3	11	41	0.74 [0.58, 0.86]	0.93 [0.81, 0.99]
Alifano 1998b	29	5	13	39	0.69 [0.53, 0.82]	0.89 [0.75, 0.96]
Banerjee 2003a	13	13	17	19	0.43 [0.25, 0.63]	0.59 [0.41, 0.76]
Banerjee 2003b	23	4	7	28	0.77 [0.58, 0.90]	0.88 [0.71, 0.96]
Caminero 1993	16	0	14	48	0.53 [0.34, 0.72]	1.00 [0.93, 1.00]
Caminero 1994	18	2	38	29	0.32 [0.20, 0.46]	0.94 [0.79, 0.99]
Chierakul 2001	31	21	36	31	0.46 [0.34, 0.59]	0.60 [0.45, 0.73]
Gevaudan 1992a	26	47	0	147	1.00 [0.87, 1.00]	0.76 [0.69, 0.82]
Gevaudan 1992b	6	9	20	185	0.23 [0.09, 0.44]	0.95 [0.91, 0.98]
Gevaudan 1992c	53	47	3	147	0.95 [0.85, 0.99]	0.76 [0.69, 0.82]
Gevaudan 1992d	18	9	38	185	0.32 [0.20, 0.46]	0.95 [0.91, 0.98]
Gevaudan 1992e	25	47	0	147	1.00 [0.86, 1.00]	0.76 [0.69, 0.82]
Gevaudan 1992f	0	9	25	185	0.00 [0.00, 0.14]	0.95 [0.91, 0.98]
Gevaudan 1992g	34	47	0	147	1.00 [0.90, 1.00]	0.76 [0.69, 0.82]
Gevaudan 1992h	0	9	34	185	0.00 [0.00, 0.10]	0.95 [0.91, 0.98]
Kunter 2003a	23	5	65	32	0.26 [0.17, 0.37]	0.86 [0.71, 0.95]
Kunter 2003b	42	3	46	34	0.48 [0.37, 0.59]	0.92 [0.78, 0.98]
Kunter 2003c	52	7	36	30	0.59 [0.48, 0.69]	0.81 [0.65, 0.92]
Luh 1996	28	17	30	207	0.48 [0.35, 0.62]	0.92 [0.88, 0.96]
McConkey 2002	27	13	29	61	0.48 [0.35, 0.62]	0.82 [0.72, 0.90]
Nsanze 1997a	18	0	17	35	0.51 [0.34, 0.69]	1.00 [0.90, 1.00]
Nsanze 1997b	4	0	31	35	0.11 [0.03, 0.27]	1.00 [0.90, 1.00]
Ratanasuwan 1997a	13	4	27	149	0.33 [0.19, 0.49]	0.97 [0.93, 0.99]
Ratanasuwan 1997b	8	4	2	151	0.80 [0.44, 0.97]	0.97 [0.94, 0.99]
Senol 2007	4	4	13	56	0.24 [0.07, 0.50]	0.93 [0.84, 0.98]

Figure 3. Sensitivity of T-SPOT.TB in HIV-negative people with confirmed active tuberculosis (modified from (24)).

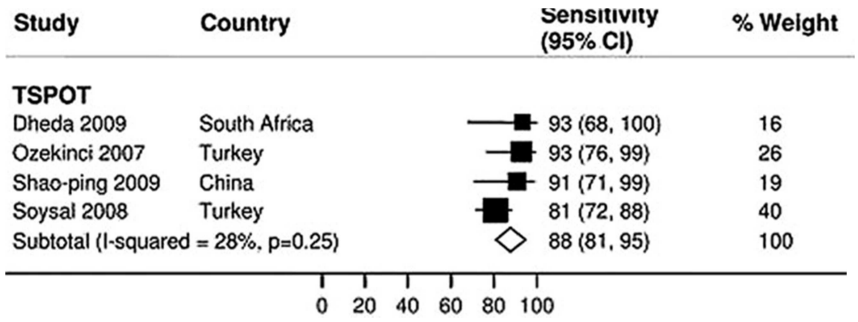


Figure 4. Example of funnel plot suggesting publication bias from (17).

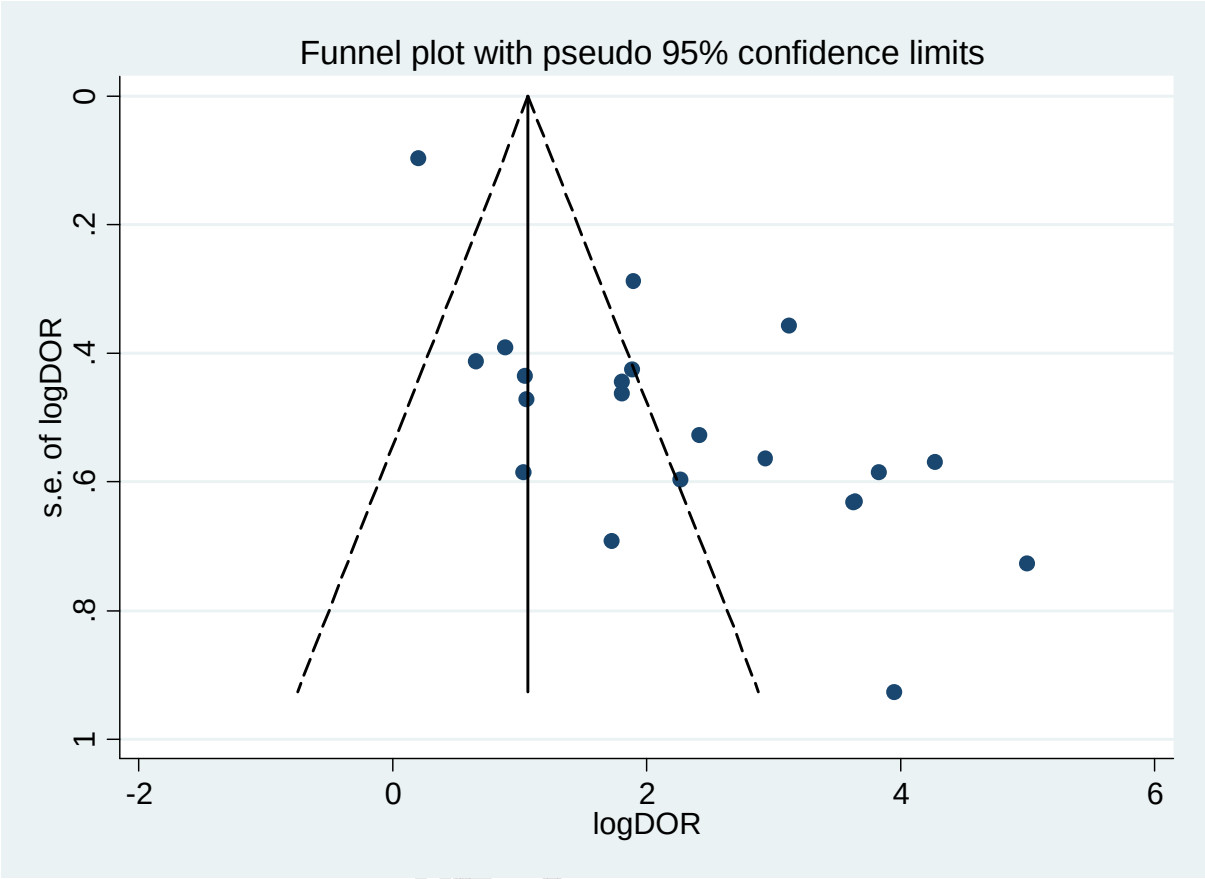
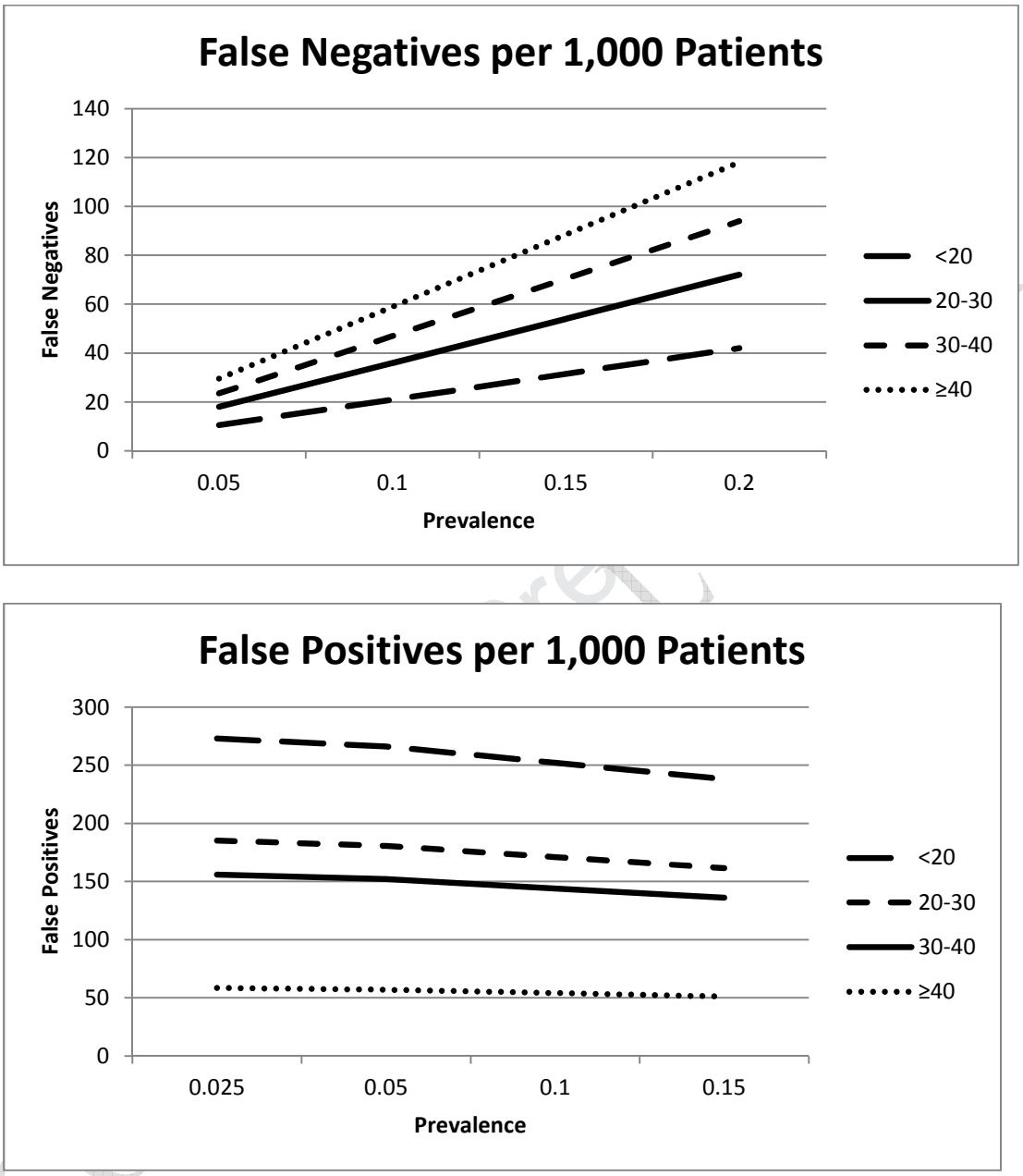


Figure 5. Example of dose response relations for different test thresholds (17)



Key findings

Rating the certainty of the body of evidence (quality of evidence or confidence in estimates) on the domains imprecision, inconsistency and publication bias for test accuracy studies shares the fundamental logic of the GRADE approach for intervention, prognostic or other studies but requires different operationalization.

What this adds to what is known?

Evidence evaluation will often begin with an evidence synthesis - ideally a systematic review or health technology assessment - and the rating of certainty in test accuracy includes assessing inconsistency, imprecision, publication bias and other domains. In this part 2 of GRADE guidance 21, we describe the judgments on these domains and across a body of evidence using examples from how GRADE has been applied to test accuracy studies in Cochrane and other reviews as well as World Health Organization and other guidelines.

What are the implications, what should change now?

Further work is needed for better operationalization of the domain imprecision and domains that may lead to increasing the certainty. However, investigators interested in using the GRADE for diagnostic and healthcare related tests should consider the guidance offered in this article for the corresponding domains and how the information is presented in evidence profiles and summary of findings tables.

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.