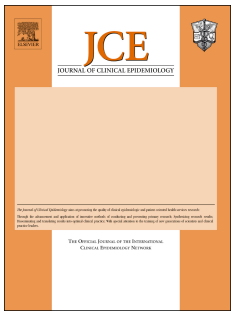
Accepted Manuscript



GRADE Guidelines: 22. The GRADE approach for tests and strategies - from test accuracy to patient important outcomes and recommendations

Holger J. Schünemann, Reem A. Mustafa, Jan Brozek, Nancy Santesso, Patrick M. Bossuyt, Karen R. Steingart, Mariska Leeflang, Stefan Lange, Tommaso Trenti, Miranda Langendam, Rob Scholten, Lotty Hooft, Mohammad Hassan Murad, Roman Jaeschke, Anne Rutjes, Jasvinder Singh, Mark Helfand, Paul Glasziou, Ingrid Arévalo Rodriguez, Elie A. Akl, Jonathan J. Deeks, Gordon H. Guyatt, GRADE Working Group

PII: S0895-4356(17)31095-8

DOI: https://doi.org/10.1016/j.jclinepi.2019.02.003

Reference: JCE 9824

To appear in: *Journal of Clinical Epidemiology*

Received Date: 11 December 2017

Revised Date: 14 November 2018

Accepted Date: 4 February 2019

Please cite this article as: Schünemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, Leeflang M, Lange S, Trenti T, Langendam M, Scholten R, Hooft L, Murad MH, Jaeschke R, Rutjes A, Singh J, Helfand M, Glasziou P, Rodriguez IA, Akl EA, Deeks JJ, Guyatt GH, GRADE Working Group, GRADE Guidelines: 22. The GRADE approach for tests and strategies - from test accuracy to patient important outcomes and recommendations, *Journal of Clinical Epidemiology* (2019), doi: https://doi.org/10.1016/j.jclinepi.2019.02.003.

GRADE Guidelines: 22. The GRADE approach for tests and strategies - from test accuracy to patient important outcomes and recommendations

Holger J Schünemann[1,2,3], Reem A. Mustafa[1,4], Jan Brozek[1,2,3], Nancy Santesso[1,3], Patrick M Bossuyt[5], Karen R Steingart[6], Mariska Leeflang[4], Stefan Lange[7], Tommaso Trenti[8], Miranda Langendam[4], Rob Scholten[9], Lotty Hooft[9], Mohammad Hassan Murad[10], Roman Jaeschke[1,2], Anne Rutjes[11], Jasvinder Singh[12], Mark Helfand[13], Paul Glasziou[14], Ingrid Arévalo Rodriguez[15], Elie A. Akl[16], Jonathan J Deeks[17], Gordon H Guyatt[1,2] for the GRADE Working Group

1. Department of Health Research Methods, Evidence, and Impact, 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. McMaster GRADE centre, 1280 Main Street West, McMaster University, Hamilton, Ontario L8S4K1, Canada
4. Department of Medicine, University of Kansas Medical Center, Kansas City, Kansas, USA
5. Clinical Epidemiology and Biostatistics and Bioinformatics Academic Medical Center, University of Amsterdam, Meibergdreef 9, Room J1b-209, P.O.Box 227001100 DE, Amsterdam, The Netherlands
6. Cochrane Infectious Diseases Group, Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK, Tel: +1 646 2439043
7. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen/Institute for Quality and Efficiency in Health Care (IQWiG), Im Mediapark 8, 50670 Köln, Germany Cologne, Germany
8. Azienda Ospedaliera Universitaria e Azienda USL di Modena, Nuovo Ospedale S. Agostino Estense, Via Giardini 1355 41126 Modena Italy
9. Cochrane Netherlands/Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508GA Utrecht, The Netherlands
10. Division of Preventive Medicine, Mayo Clinic, 200 1st ST. SW, Rochester, MN, 55902
11. Clinical Trial Unit (CTU) Bern, Institute of Primary Health Care; Institute of Social and Preventive Medicine, University of Bern, Switzerland
12. Medicine Service, VA Medical Center, Birmingham, AL, and Department of Medicine, University of Alabama at Birmingham, 510, 20th Street South, FOT805B, Birmingham, AL, USA
13. Oregon Evidence-based Practice Center, Oregon Health & Science University, Portland VA Medical Center, Portland, Oregon
14. CREBP, Faculty Health Science & Medicine, Bond University, Gold Coast, Qld 4229
15. 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.
16. Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

17. Department of Internal Medicine, American University of Beirut, Riad-El-Solh Beirut, Beirut 1107 2020, Lebanon.

Email addresses:
Holger J Schünemann <schuneh@mcmaster.ca>; Reem Mustafa <ramustafa@gmail.com>; Brozek, Jan <brozekj@mcmaster.ca>; Nancy Santesso <santesna@mcmaster.ca>; Patrick Bossuyt <p.m.bossuyt@amc.uva.nl>; Karen Steingart < karenst@uw.edu >; Mariska Leeflang <m.m.leeflang@amc.uva.nl>; Stefan Lange <stefan.lange@iqwig.de>; Tommaso Trenti <t.trenti@ausl.mo.it>; Rob Scholten < R.J.P.Scholten@umcutrecht.nl >; Jasvinder Singh <Jasvinder.Singh@ccc.uab.edu>; Miranda Langendam <m.w.langendam@amc.uva.nl>; Lotty Hooft < L.Hooft@umcutrecht.nl>; Mohammed Hassan Murad <Murad.mohammad@mayo.edu>; Roman Jaeschke <jaeschke@mcmaster.ca>; Rutjes <anne.rutjes@ispm.unibe.ch>; Mark Helfand <helfand@ohsu.edu>; Paul Glasziou < Paul Glasziou <pglaszio@bond.edu.au>; Ingrid Arévalo Rodriguez <inarev7@yahoo.com>; Jon Deeks j.deeks@bham.ac.uk; Elie A Akl <ea32@aub.edu.lb>; Gordon Guyatt <guyatt@mcmaster.ca>;

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Address for correspondence:
Holger J. Schünemann, MD, PhD, MSc
Chair, Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
Department of Clinical Epidemiology and Biostatistics
McMaster University, HSC Room 2C16
1280 Main Street West Hamilton, ON L8S 4K1
Canada
Tel: +1 905-525-9140 x 24931
Email: schuneh@mcmaster.ca

Word count: 4367

Boxes: 1

Tables: 3

Figures: 5

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Highlights

GRADE has developed and applied a comprehensive framework to evaluate the certainty of the evidence when direct evidence of the effect of testing on outcomes is not available and linked evidence that connects test accuracy to downstream consequences is required for decision-making. Ideally, this linked evidence comes from systematic reviews is that informs analytical frameworks for questions related to tests.

What this adds to what is known?

Application of GRADE's approach requires guideline developers to rate their certainty in each element of the linked evidence that is required for decision-making.

What are the implications, what should change now?

Further research should address ways to arrive at, for each critical or important outcome and across outcomes, ratings of the certainty derived from the linked sources of evidence. This will often require formal modelling and assessing the certainty in the models.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Abstract:

Objectives: This article describes GRADE's framework of moving from test accuracy to patient or population important outcomes. We focus on the common scenario when studies directly evaluating the effect of diagnostic and other tests or strategies on health outcomes are not available or are not providing the best available evidence.

Study Design and Setting: Using practical examples, we explored how guideline developers and other decision makers can use information from test accuracy to develop a recommendation by linking evidence that addresses downstream consequences. Guideline panels should develop an analytic framework that summarizes the actions that follow from applying a test, and the consequences.

Results: We describe GRADE's current thinking about the overall certainty of the evidence (also known as quality of the evidence or confidence in the estimates) arising from consideration of the often complex pathways that involve multiple tests and management options. Each link in the evidence can – and often does - lower the overall certainty of the evidence required to formulate recommendations and make decisions about tests. The frequency with which an outcome occurs and its importance will influence whether or not a particular step in the linked evidence is critical to decision-making.

Conclusions: Overall certainty may be expressed by the weakest critical step in the linked evidence. The linked approach to addressing optimal testing will often require the use of decision analytic approaches. We present an example that involves decision modeling in a GRADE Evidence to Decision framework for cervical cancer screening. However, since resources and time of guideline developers may be limited, we describe alternative, pragmatic strategies for developing recommendations addressing test use.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

GRADE Guidelines: 22. The GRADE approach for tests and strategies - from test accuracy to patient important outcomes and recommendations

1. Introduction

Two previous articles in this series describe how systematic review authors and guideline developers develop GRADE Evidence to Decision Frameworks and assess their certainty of a body of evidence (also referred to as confidence in estimates or quality of evidence) evaluating tests, test strategies, management and downstream consequences.(1)(add reference to article 1) We focused on applying GRADE to accuracy studies but did not describe how different bodies of evidence are linked and how to rate the overall certainty of the evidence. Prior work on this topic has dealt with issues such as causal pathways in answering questions about tests, analytic frameworks and the importance of considering health outcomes for decision-making.(2-7) In the present article we focus on GRADE's operationalization of linking information from accuracy studies to important outcomes when investigations directly evaluating the effect of diagnostic or other tests and related strategies on important outcome are unavailable. We will describe the conceptual issues of how guideline developers can move from the linked evidence that connects test accuracy to downstream management and consequences, and certainty of that evidence, to recommendations regarding the use of diagnostic and other tests use for the purpose of screening, monitoring and surveillance. With this work we are expanding on GRADE's approach to rating the certainty of evidence for questions about tests.(5)

The first part of the article will describe the judgments about the certainty we can place in the linked evidence between test accuracy and important outcomes. GRADE considers

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

accuracy a surrogate marker that requires further evaluation of the related consequences, including management decisions and consequent health outcomes. In particular, our framework will show why guideline panels and other decision makers should usually be cautious when they use evidence of test accuracy as the basis for recommendations or decision. This caution is necessary because decision-making requires a review of and judgements about the evidence that links the evidence regarding accuracy to patient or population-important outcomes.

The second part of the article focuses on the criteria that are involved in moving from evidence to a recommendation or decision using examples from guidelines that have applied this approach to diagnostic tests and strategies.(1) These examples, which expand on explanations previously provided regarding EtD frameworks, make clear that optimal insight into the consequences of alternative diagnostic strategies will often require full decision analytic modelling.(1)

Decision analytic modelling, often undertaken when evidence is limited and use of expert estimates is necessary, involves constructing a decision tree and then forecasting based on probabilities of possible outcomes. The objective of a decision analysis is to discover the most advantageous alternative under the defined conditions. We will provide an example of using decision analysis to model the relation between test accuracy, treatment strategies and health outcomes.

In many situations, however, limited guideline resources and time mandate other, less sophisticated approaches. Less sophisticated approaches are also desirable when the

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

correct decision or recommendation is evident without a full decision analysis. Therefore, we will present examples of less resource-intensive pragmatic approaches. We will conclude by summarizing challenges and suggestions for future work.

2.0 What evidence is needed to make deductions about effects on patient outcomes?

The GRADE approach to recommendations regarding test use involves balancing the desirable and the undesirable consequences (including non-health related consequences such as resource utilization and impact on equity).(1, 8-10) As they apply GRADE to recommendations about tests, guideline panelists will recognize that a division between testing and therapy or treatment is artificial. Figure 1 highlights that testing, including for diagnosis, screening and monitoring, demands interventions that become part of the overall strategy to management and that have consequences that require consideration.(5) The interventions that follow from test results may include observation of people or patients when no further action is required or possible.

Guideline panels should develop an analytical framework that clarifies the interventions that follow from applying a test, and their consequences. (1, 7) Figure 2 provides a simplified and generic example of such a framework applied to a screening intervention.(11) The ideal body of evidence would include studies that directly compare the test strategies under consideration (i.e. randomized trials) and the resulting interventions and consequences (i.e. patient-important outcomes). Such studies would, by design, address all of the issues in the analytical framework, and allow guideline panelists to apply the familiar GRADE approach for interventions articulated in detail in previous articles in this series. For most tests or test and treat strategies, however, this direct evidence does not exist.(12)

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Figure 3 describes the analytical framework including the health care pathways from a World Health Organization guideline addressing screening and treatment of cervical intraepithelial neoplasia (CIN stage 2 or 3), a precursor for cervical cancer, in low and middle resource countries (Box 1, example 1). Using a detailed and structured process, the guideline panel, comprising experts from various disciplines, considered two screening options: human papilloma virus detection (HPV) or visual inspection with acetic acid (VIA), and the possible subsequent treatment options.(13-15)

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Box 1.

Example 1: Guideline determining population important outcomes and modeling to estimate benefits and harms

In women at risk for cervical intraepithelial neoplasia *(CIN) stage 2 to 3 in low and middle-income settings, what is the impact of testing for presence of HPV instead of VIA followed by management on patient and population important outcomes?(13)*

Population: women at risk of cervical cancer in low and middle-income countries
Intervention: one time screening and treatment for CIN 2-3
Comparison: no screening and treatment program
Purpose and role of test: diagnosis and replacement
Health outcomes: modeled estimates of death from cervical cancer, cervical cancer incidence, CIN 2-3 recurrence, major bleeding, premature delivery, infertility, major and minor infections, unnecessary treatment, cervical cancer detection during screening

Recommendations: Where resources permit, the expert panel suggests a strategy of screen with an HPV test and treat with cryotherapy (or LEEP when not eligible for cryotherapy) over a strategy of screen with VIA and treat with cryotherapy (or LEEP when not eligible) (conditional recommendation, ⊕⊖⊖⊖ very low certainty evidence)
In resource-constrained settings, where screening with an HPV test is not feasible, the expert panel suggests a strategy of screen with VIA and treat with cryotherapy (or LEEP when not eligible) over a strategy of screen with an HPV test and treat with cryotherapy (or LEEP when not eligible) (conditional recommendation, ⊕⊖⊖⊖ very low certainty evidence).
Rationale: The benefits of screen-and-treat with an HPV test or VIA, compared to no screening, outweighed the harms, but the reductions in cancer and related mortality were greater with an HPV test when compared to VIA. The availability of HPV testing is resource-dependent and, therefore, the expert panel suggests that an HPV test over VIA be provided where it is available, affordable, implementable, and sustainable over time.

Example 2: Guideline using detailed case scenarios and simple modeling

In patients suspected of cow's milk allergy (CMA), what is the impact of using skin prick tests rather than an oral food challenge with cow's milk on the diagnosis and management of IgE-mediated CMA?(16, 17)

Population: patients suspected of CMA
Intervention: skin prick test
Comparison: no skin prick test before or after oral food challenge
Purpose and role of test: diagnosis and replacement or add on
Health outcomes: described as scenarios. For example, consequences of true positives were described as "The child will undergo oral food challenge that will turn out positive with risk of anaphylaxis, albeit in controlled environment; burden on time and anxiety for family; exclusion of milk and use of special formula. Some children with high pretest probability of disease and/or at high risk of anaphylactic shock during the challenge will not undergo challenge test and be treated with the same consequences of treatment as those who underwent food challenge.
Recommendation: In settings where oral food challenge is considered a requirement for making a diagnosis of IgE-mediated CMA, we recommend using oral food challenge with cow's milk as the only test without performing a skin prick test as a triage or an add-on test to establish a diagnosis (strong recommendation, low certainty of the evidence).
Rationale: No additional benefit for patient outcomes, no need to do other testing if the reference test is performed regardless.

Example 3: Guideline using case example for modeling and considering the possible consequences explicitly but no case scenarios or impact on health outcomes modeled

In patients suspected of having pulmonary tuberculosis (TB), what is the impact of commercial serological tests for TB rather than conventional tests such as smear microscopy?(18)

Population: patients suspected of having pulmonary tuberculosis
Intervention: serological testing
Comparison: sputum microscopy
Purpose and role of test: diagnosis and replacement
Health outcomes: a case study linking the test accuracy data to active tuberculosis in India showed that replacing sputum microscopy with serological testing would result in an estimated 14,000 additional cases of TB diagnosed but also result in 121,000 additional false-positive diagnoses relative to microscopy.

Recommendation: … recommended that these tests should not be used in individuals suspected of active pulmonary or extra-pulmonary TB (strong recommendation).
Rationale: The panel considered the consequences of false positives as unacceptable in balance with the possible benefits of detecting additional cases of TB. The overall harms far outweigh any potential benefits.

For instance, only some lesions are amenable to cryotherapy. If lesions are not, possible therapeutic interventions include cold knife conization or loop electrosurgical excision

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

procedure (LEEP). The panel considered the possible outcomes – including benefits and harms - likely to result from each of the possible screen and treat pathways.

Figure 4 describes the sequential steps that are important when assessing the certainty of evidence following the consequences of testing and treating. Having formulated the question focusing on patient important outcomes and having concluded that direct evidence addressing these outcomes does not exist, the approach begins with an assessment of test accuracy (step 1). It proceeds to assessing the important consequences that may arise from applying the competing tests or diagnostic strategies and the evidence that links test results to those consequences (step 2).

Returning to the example of the cervical screening guidelines, in which direct evidence for patient important outcomes is unavailable, step 1 included a systematic review of the body of evidence describing the performance of the two screening tests against a reference standard. The authors of the systematic review conducted a meta-analysis to obtain summary estimates of the sensitivity and specificity of the two screening tests.(15) The resulting pooled sensitivity of 95% (95% CI: 84 to 98) and pooled specificity 84% (95% CI: 72 to 91) for HPV, and pooled sensitivity of 69% (95% CI: 54 to 81) and a pooled specificity of 87% (95% CI: 79 to 92) for VIA, was based on five studies that compared both tests against a reference standard.

Applying these summary statistics to the best estimates of the pretest probability of the target population (assumed here to be 2%) informed estimates of the number of test positives (true positives and false positives) and test negatives (true negatives and false

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

negatives). For example, 1.9% (19 per 1000 women or 95% of 2%) in the HPV and 1.4% (14 per 1000 women or 69% of 2%) in the VIA group would be true positives. Although HPV's accuracy is clearly superior (much better sensitivity with similar specificity) the considerably greater cost of HPV required estimating the magnitude of effect of alternative test strategies on patient-important outcomes (i.e. are the benefits at varying levels of pre-test probability worth the costs, in particular for country wide screening programs in low and middle-income countries).

Step 2 involved linking these test outcomes to the anticipated important outcomes. The panel used both a literature review and the experience of a multidisciplinary panel to provide information about the outcomes women may experience. Women with a positive test, suggesting the presence of CIN 2-3, would undergo further management with one of the possible therapies to reduce the risk of cervical cancer. Each treatment is associated with a probability of cure and a probability of adverse effects. Those undergoing the procedures would, however, include not only those with CIN 2-3 (true positives) but also women without CIN 2-3 (false positives), and both groups would experience the adverse consequences, the latter without experiencing the benefits.

Women with a negative test result, including those with a false negative test result and the attendant risk of CIN 2-3 developing into cervical cancer, would undergo only further observation. Although this model ignores the possibility of repeating the screening test, in some settings - such as low and middle-resource countries - women may undergo only a single screening test and not return for further testing. This makes the model suitable for those situations in which women are not repeatedly screened.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Estimates of treatment effects, both beneficial and harmful, and natural history should, ideally, come from systematic reviews of the relevant evidence.(13) For example, just as the efficacy of cryotherapy should be evaluated with a systematic review, so should the risk of developing cervical cancer in untreated CIN 2-3 (the natural history of the disease that determines the outcomes of false negatives). The systematic review determining the efficacy of cryotherapy reported a 61% relative risk reduction based on observational data (19) and the evidence addressing the natural history suggested a 2% progression to cervical cancer over 30 years.(14, 20)

Turning to a different example (Box 1, example 2), for a recommendation addressing the use of skin prick testing for cows' milk allergy (CMA), a condition affecting between 2 and 5% of children, the guideline panel evaluated the possible benefits and downsides on the basis of case examples using semi-quantitative information. (21, 22) For instance, in order to understand the consequences associated with the 264 per 1000 false negative skin prick tests in a population with a high pretest probability of CMA, guideline panel members received and reviewed typical case scenarios:

i) The child suspected of CMA will be allowed to return home and will have an allergic reaction (possibly anaphylactic) to cow's milk at home; high parental anxiety and reluctance to introduce future foods following the reaction; may lead to multiple exclusion diet.

ii) The real cause of symptoms (i.e. CMA) will be missed leading to unnecessary investigations and treatments.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

The panel developed these case scenarios, the baseline risk and the possible consequences, on the basis of a review of the literature and information obtained from allergists with experience in caring for affected patients.

In another guideline (Box 1, example 3), a WHO guideline panel considered the consequences of applying a commercial serological tests sensitivity of 59% and specificity of 91% in a population with a 10% risk of pulmonary tuberculosis: the test results in 41 per 1000 false negatives and 81 per 1000 false positives. (18) Guideline panel members applied existing evidence synthesized in tuberculosis treatment guidelines to link the treatment efficacy and possible detrimental effects from delayed diagnosis, confusing other respiratory diseases (such as pneumonia) with pulmonary TB and resulting death from another disease, adverse drug reactions and unnecessary consumption of health care and patient resources. Panel members also reviewed a model describing the results in terms of true and false positive and true and false negatives of serological testing compared against other TB testing modalities (sputum smear and culture), including sensitivity analyses. The information facilitated understanding of the effects in terms of unnecessarily treated and stigmatized patients and those appropriately treated.

3.0 How should guideline panels rate the certainty of the evidence on health outcomes

Preferably, guideline developers will evaluate and rate a body of evidence for each component of the linked evidence required for decision-making. Ideally, the evidence synthesis will be based on a systematic review (Figure 5).

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

3.1 Rating certainty in the estimates of test accuracy

As described in the prior article, when direct evidence about important health outcomes is not available or associated with low certainty, GRADE begins by assessing the certainty of the estimates of the accuracy of the competing diagnostic strategies.(reference to article 1) We also clarified the roles of tests and how to interpret accuracy in the comparison of a new against an existing test. The systematic review of HPV and VIA reported high certainty in the test sensitivity, but important inconsistency in the specificity estimates across the 5 eligible studies yielding an overall certainty rating of moderate for specificity (Table 1).(15) In previous articles we also introduced the three layers of SoF Tables for questions about the use of tests.(add reference to article dx 1)(23) Layer 1 and 2 SoF Tables , described in the previous article, do not consider the directness of the relation between test accuracy and health outcomes; here we focus on the third layer that considers the directness of the relation between test accuracy and health outcomes.

3.2 Rating the certainty in linked evidence – directness of the health outcomes

Here, we consider the assessment of the certainty of the evidence for a question involving tests for which direct evidence of effect on important outcomes is not available. In an ideal situation, rating the body of evidence should consider, for each important or critical outcome, the certainty of evidence from each linked element. In other words, assessing the linked evidence completes the assessment of how directly test accuracy estimates relate to the outcomes and informs the ratings of GRADE's indirectness domain.

For example, uncertainty regarding the estimates of the pretest probability or baseline risk used to calculate the test results in Table 1 will influence the overall rating of the certainty

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

(Figure 4, step 1). Application of GRADE for prognostic studies or prevalence studies will inform this rating of certainty.(24, 25) Similarly the certainty of the evidence regarding the estimates of the treatment effects of cryotherapy and other treatments should influence the overall certainty. Applying GRADE for interventions, the certainty in the estimates for the effects of cryotherapy - coming from observational studies with high risk of bias - was very low (Figure 4, step 2).(13) Persistence of CIN 2-3 in false negatives was estimated as approximately 70% based on moderate certainty evidence from longitudinal prognostic observational studies. Thus, step 2 in Figure 4 involves a rating of the certainty in the estimates when going from the test results to important outcomes. Because the authors of the cervical cancer guideline had very low certainty in some of the steps that were part of the linked bodies of evidence, they rated the overall certainty as very low.(14)

In addition, the rating of directness must include an assessment of the degree to which the impact of the test on patients occurs solely through the consequences of classification as true negative, false positive, true positive or false negative. Tests may impact on patient outcomes where they are less invasive, can be delivered in different settings, care pathways or at different time points in the course of disease, or where they provide additional information which can change adherence or the nature of the intervention given (1, 5-7). For example, point of care testing that have the potential to speed up decision-making, enabling patients to access appropriate treatment faster and reduce anxiety waiting for results. A linked evidence evaluation that solely considers the consequences of treatment based on test results will not capture the benefits of using a quicker and more portable test, and potentially will be misleading (1, 5, 26).

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Table 2 presents the layer 3 summary of findings (SoF) Table for tests based on the best available research evidence (layer 2 SoF Tables add direct estimates of the adverse consequences and inconclusive results from tests and certainty of those estimates to layer 1 tables but is not provided here) and a link to an interactive summary of findings table in GRADE's database of evidence profiles and evidence to decision frameworks (dbep.gradepro.org; also in the online appendix). Certainty estimates do not appear in the table 2 (layer 3) because their presentation would be inefficient (they are all very low). The explanations in the related text of the table provide the sources of evidence, assumptions made, and explanations. Table 3 summarizes the different layers of GRADE SoF tables, what information each of them includes and who would develop or use the tables.

For example, the guideline panel assumed mortality will decrease in true positives relative to those untreated due to treatment, and will increase in false negative due to late or missed diagnosis. It also assumed that there would be no mortality from cervical cancer in true negative and false positive; this, however, is a simplification because women may develop CIN 2-3 later in life. The decision analysis, based on the available evidence, applied a 70% natural persistence of CIN 2-3 and progression to cervical cancer in women who do not receive treatment.(20)

Assumptions of the recurrence rates of CIN 2-3 were 5.3% after successful initial cryotherapy, 2.2% in CKC and 5.3% in LEEP. That is, for a pretest probability of 2% or 20,000 for CIN 2-3 out of 1 million, when using human papilloma virus (HPV) test for diagnosis of

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

CIN 2-3 (pooled sensitivity 95% and pooled specificity 84%), there would be 19,000 TP and 1,000 FN. If all of the 19,000 TP receive treatment and assuming 2.2% recurrence or persistence rate with CKC treatment, 18,582, out of the 19,000 women screened would be cured but 418 (2.2% of 19,000) would relapse with CIN2-3.

On the other hand, the 1,000 FN will not receive treatment. Assuming 70% natural persistence of CIN 2-3 (30% natural regression), CIN 2-3 will persist in 700 out of 1,000 FN. Based on the available observational data, approximately 2.5% of women who were not cured would progress to cervical cancer over a year. So, 2.5% out of 1118 (700 from FN + 418 from TP) or 28 women will progress to cervical cancer. Then, out of the 19,000 TP and 1,000 FN, a total of 1090 (418 + 700 - 28) recur or persist as CIN 2-3.

Based on the available observational data, approximately 2.5% (350 out of 14,000 women with persistent CIN 2-3 with no treatment out of 1 million women screened) of those who were not cured would progress to cervical cancer over a year. So out of 1090, 28 women will develop cervical cancer. Based on a 71% mortality rate, 20 out of the 28 will die from cervical cancer. Other evidence suggested that no major bleeding would occur in the test negatives (both TN and FN) as they were not treated, but a small proportion of women (0.03%) would experience a major bleed with cryotherapy, as would 0.9% of women treated with CKC based on reported proportions in single arm studies.

This modeling was conducted for each of the pathways in Figure 3 and results finally aggregated in a modified layer 3 SoF Table (Table 2). Table 2 summarizes all effects by outcome for each test-screen strategy to allow balancing the benefits and harms. In

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

situations when circumstances change and a health status determines subsequent outcomes, more complex modelling will be beneficial to guideline developers and users.

4. How do the certainty ratings of the linked evidence influence the overall rating of the certainty of the evidence

Having realized that each link in the evidence can – and often does - lower the overall certainty one has in the evidence for each outcome, there are two viable options for describing an overall rating of the certainty of the evidence for each outcome.

Option 1. Evaluate which bodies of linked evidence are critical for decision-making and base the overall rating of the certainty for population important outcomes on the lowest certainty of these bodies of evidence. For example, if a panel decided that both the natural history of the disease and efficacy of cryotherapy were critical for the decision, despite high certainty of the evidence of the estimates of test accuracy for TP and FN and moderate confidence in TN and FP, the recommendation would be associated with a rating of very low certainty for each of the outcomes of interest. The frequency and importance with which an outcome occurs will determine whether or not linked evidence is critical to decision-making. This is the approach the guideline panel on cervical cancer screening and treatment took by rating the overall certainty as very low. An advantage of this approach is that it is most consistent with GRADE's focus on outcomes that are critical to decision-making.

Option 2. Present the evidence from test accuracy and linked evidence separately without assigning an overall rating of certainty. Using this approach, the certainty of the evidence for all elements would be described separately. For example, in the cervical cancer

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

screening and treatment guidelines, the recommendation would be accompanied by a rating of the certainty for the test accuracy (moderate certainty), the prognostic evidence (very low certainty) and the effects of therapy (e.g. very to low certainty for cryotherapy) without an overall rating of the certainty.

Despite being more complex, we suggest using option 1 as preferred approach: providing an overall rating in the certainty across the elements of the linked evidence critical for decision-making. Usually and justifiably, certainty of the evidence that results from linking test accuracy evidence with people important outcomes will result in low or very low certainty of the evidence. This is the reason why guideline panels and other decision makers should be cautious when they use evidence of test accuracy as the basis for recommendations or decision. Further work will inform how users of GRADE's summaries of a body of linked evidence and recommendations will appropriately integrate ratings of certainty and how to improve this presentation.

5. When is evidence about accuracy sufficient to make a decision?

Comprehensively (re)evaluating all evidence that informs the consequences of testing requires time and resources. Informed decisions require, however, transparent presentations of the considerations influencing the recommendations and the certainty in the underlying evidence.

If the accuracy of one test is very similar or superior to another test and there is no shift in the type of people or patients classified by the test, and it is also less expensive, easier to use, or comes with less harm, guideline panels can confidently recommend the test without

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

explicit consideration of downstream consequences. For example, the use of a shortened version of a psychological test, e.g. mini mental exam, that has fewer items but the same accuracy would save time without adverse consequences. Using computed tomography or magnetic resonance for presumed cerebrospinal fluid obstruction or lesions of the posterior fossa instead of an air encephalogram, a technique used before the arrival of modern imaging, provides greater accuracy and fewer risks. In such situations, panels can omit modeling the effects on health outcomes and impact on other consequences of different accuracy. This issue is addressed elsewhere in detail.(27)

6. How can panels make recommendations about tests or care strategies involving tests

A recommendation associated with a diagnostic question follows from an evaluation of the balance between the desirable and undesirable consequences of the test and subsequent management (Figure 1).(1) When estimates of the consequences of the false positive, false negative, inconclusive results and complication rates with the alternative diagnostic strategies warrant high certainty, we can make strong inferences concerning the relative effects of a test on important outcomes. Such situations are, however, rare; typically guideline panels will lower their overall certainty in the evidence when considering linked evidence.

Box 1 describes three recommendations regarding the use of tests that rely on test accuracy and the approaches the guideline panel took to develop them. These three examples emphasize that careful consideration of the consequences, and ideally decision analysis, are required to develop recommendations.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

The guideline panel that developed recommendations regarding skin prick testing in patients with cow milk allergy determined that for patients with a relatively low probability of the disease (approximately 10%) skin prick testing results in a large number of false positives leading to unnecessary anxiety and further testing. It also leads to missing about 3% (33/1000 tested patients are false negatives) of patients who suffer from cow milk allergy with the risk of severe allergic reaction and death. The certainty in the estimates of accuracy was low because of risk of bias and unexplained inconsistency in the 23 evaluated studies. Furthermore, the panel was uncertain about the links between tests results and patient outcomes and thus characterized the overall certainty as very low. The guideline panel making recommendations about serological testing for tuberculosis used pragmatic and quick modeling (sometimes characterized as "back of the envelope"), but still made its assumptions about consequences of test results transparent.

6.1 GRADE Evidence to Decision frameworks

GRADE has used decision tables that increase transparency of the decision-making process.(10, 28) Extensive work has informed the selection of criteria that influence the development of health care recommendations about tests. As evidence to decision (EtD) frameworks, these criteria have been further developed as part of the DECIDE project and are included in GRADE's GRADEpro software (www.gradepro.org) together with the interactive SoF tables. (1, 29-31) We have described these frameworks in detail in the 16th article in this series (1).

The purpose of the frameworks is to help guideline panels and other decision makers developing recommendations about the use of tests to move from evidence to

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

recommendations. The frameworks inform decision makers' judgments about the desirable and undesirable consequences of the considered. The frameworks also ensure that panels consider important factors that determine a recommendation (criteria) by providing a concise summary of the best available evidence. They facilitate a structured discussion and identify reasons for agreement and disagreement.

EtD frameworks should include one or more of the three layers of a SoF table for tests – and in particular a description of the expected health outcomes, ideally in a layer 3 SoF table (Table 2). An alternative is a narrative summary, but whatever the nature of the summary it should be included and linked to the full GRADE evidence profile. The framework, or the background information, should include the decision analysis (Table 2). Other information listed in Table 2 can be included when guideline panels intend to achieve complete transparency about the recommendations they make.

7. Conclusions

GRADE has developed and applied a comprehensive framework for rating the certainty in test accuracy estimates from a body of evidence and linking this evidence to outcomes when studies directly evaluating the effects of testing on health outcomes are not available. Guideline panels and other decision makers should be cautious when they use evidence of test accuracy as the basis for recommendations or decision. The framework focuses on explicitly and transparently laying out the elements or bodies of evidence required to making the link. Although the framework has facilitated the development of recommendations about tests for several guidelines (18, 21, 22, 32) and can be ready applied, further examples and future research in several areas addressing the assessment of

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

the certainty and the degree of modeling required will be necessary to move this field forward.

Disclosure Statement

The authors are members of the GRADE Working Group. JAS has received research grants from Takeda and Savient and consultant fees from Savient, Takeda, Regeneron, Merz, Iroko, Bioiberica, Crealta/Horizon and Allergan pharmaceuticals, WebMD, UBM LLC and the American College of Rheumatology. JAS serves as the principal investigator for an investigator-initiated study funded by Horizon pharmaceuticals through a grant to DINORA, Inc., a 501 (c)(3) entity. JS is a member of the executive of OMERACT, an organization that develops outcome measures in rheumatology and receives arms-length funding from 36 companies; a member of the American College of Rheumatology's (ACR) Annual Meeting Planning Committee (AMPC); Chair of the ACR Meet-the-Professor, Workshop and Study Group Subcommittee; and a member of the Veterans Affairs Rheumatology Field Advisory Committee. JS is the editor and the Director of the UAB Cochrane Musculoskeletal Group Satellite Center on Network Meta-analysis. The other authors declared no financial conflict of interest.

Acknowledgment

This work was partially funded by a European Community's Sixth Framework Programme (FP6/2001-2006) “The human factor, mobility and Marie Curie Actions Scientist Reintegration” IGR 42192 – (“GRADE” to Dr. Schünemann), the European Commission's Seventh Framework Programme (FP7/2007-2013) under grant agreement °258583 (DECIDE project), and the German Insurance Fund. Sole responsibility lies with the authors and the European Commission is not responsible for any use that may be made of the information contained therein. The systematic reviews describing the HPV/VIA example were supported by grants to HJS and NS but the work presented here reflects the interpretation of the authors not of the World Health Organization. We would like to thank the many individuals and organizations who have contributed to the progress of the GRADE approach through funding of meetings and feedback on the work described in this article. Our thanks go to

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Andrew D. Oxman who contributed significantly to the early discussions and the shaping of the approach.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

References

1. 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.
2. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. American journal of preventive medicine. 2001;20(3 Suppl):21-35.
3. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991;11(2):88-94.
4. 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.
5. 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.
6. Schunemann HJ, Mustafa R, Brozek J. [Diagnostic accuracy and linked evidence--testing the chain]. Z Evid Fortbild Qual Gesundhwes. 2012;106(3):153-60.
7. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. Bmj. 2012;344:e686.
8. 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.
9. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-25.
10. Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66(7):726-35.
11. Barton MB, Miller T, Wolff T, Petitti D, LeFevre M, Sawaya G, et al. How to read the new recommendation statement: methods update from the U.S. Preventive Services Task Force. Annals of internal medicine. 2007;147(2):123-7.
12. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol. 2012;65(3):282-7.
13. 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.
14. 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.
15. Mustafa RA, Santesso N, Khatib R, Mustafa AA, Wiercioch W, Kehar R, et al. Systematic reviews and meta-analyses of the accuracy of HPV tests, visual inspection with acetic acid, cytology, and colposcopy. International journal of gynaecology and obstetrics:

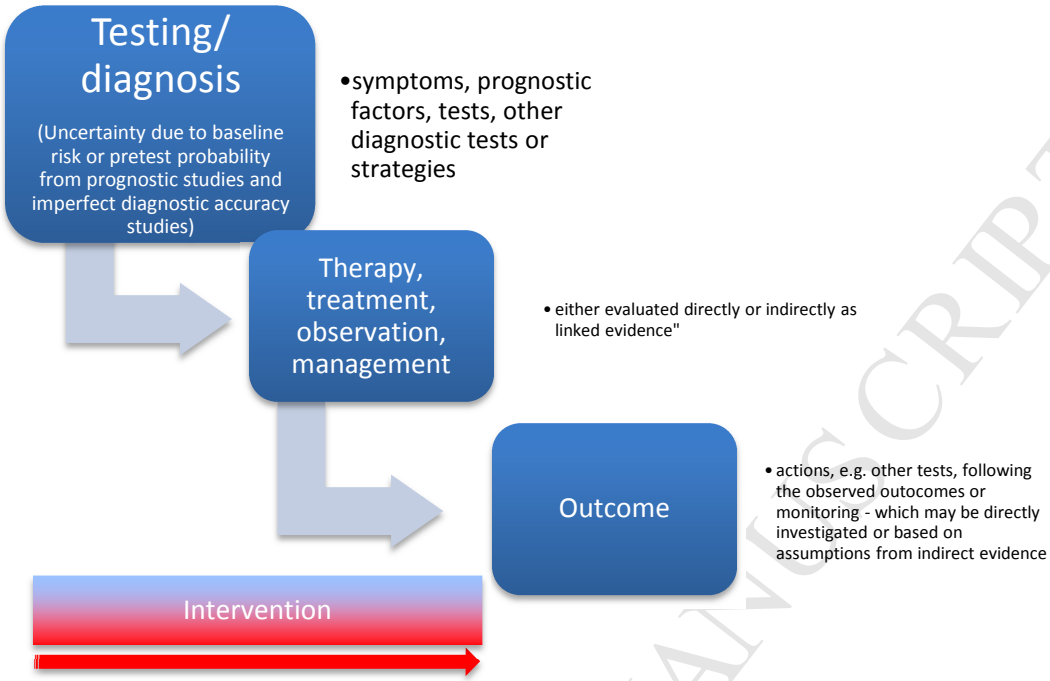
C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

the official organ of the International Federation of Gynaecology and Obstetrics. 2016;132(3):259-65.
16. 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.
17. Fiocchi A, Schunemann HJ, Brozek J, Restani P, Beyer K, Troncone R, et al. Diagnosis and Rationale for Action Against Cow's Milk Allergy (DRACMA): a summary report. J Allergy Clin Immunol. 2010;126(6):1119-28.e12.
18. WHO. Commercial Serodiagnostic Tests for Diagnosis of Tuberculosis 2011;ISBN 978 92 4 150205 4
19. Santesso N, Schunemann H, Blumenthal P, De Vuyst H, Gage J, Garcia F, et al. World Health Organization Guidelines: Use of cryotherapy for cervical intraepithelial neoplasia. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics. 2012;118(2):97-102.
20. Organization WH. [Available from: (http://globocan.iarc.fr/factsheets/cancers/cervix.asp.
21. Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: Making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implement Sci. 2011;6:62.
22. 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.
23. 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.
24. 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.
25. 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.
26. Breheny K SA, Deeks JJ. Model based economic evaluations of diagnostic point of care tests were rarely fit for purpose. J Clin Epidemiol. in press.
27. 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.
28. 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.
29. Treweek S, Oxman AD, Alderson P, Bossuyt PM, Brandt L, Brozek J, et al. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results. Implementation science : IS. 2013;8:6.
30. 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.

31. 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.
32. Bates SM, Jaeschke R, Stevens SM, Goodacre S, Wells PS, Stevenson MD, et al. Diagnosis of DVT: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e351S-418S.

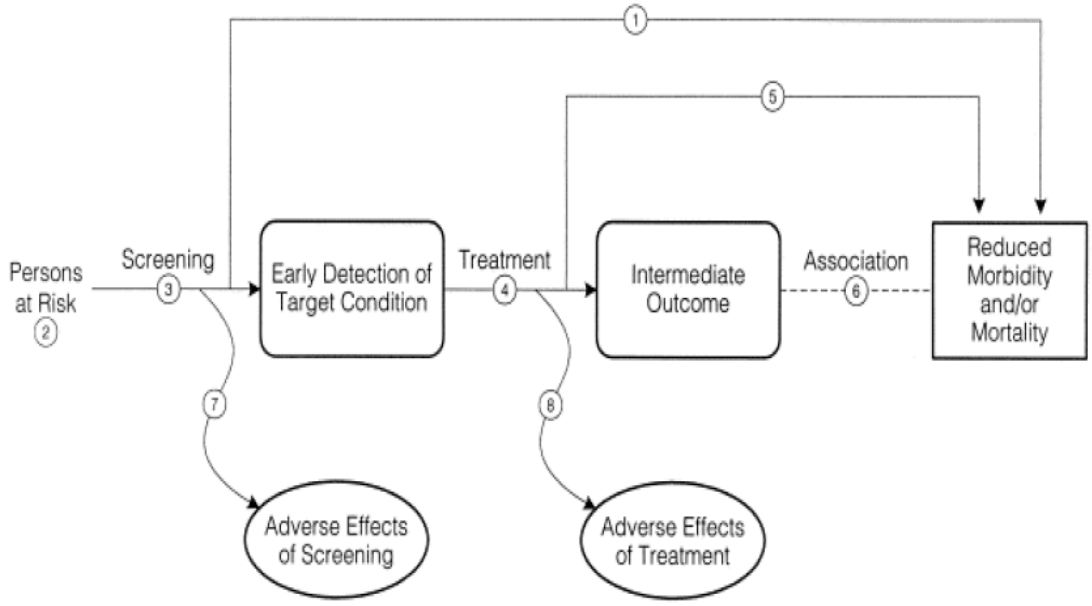
C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Figure 1. Linkage of testing, interventions and outcomes (6)



C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

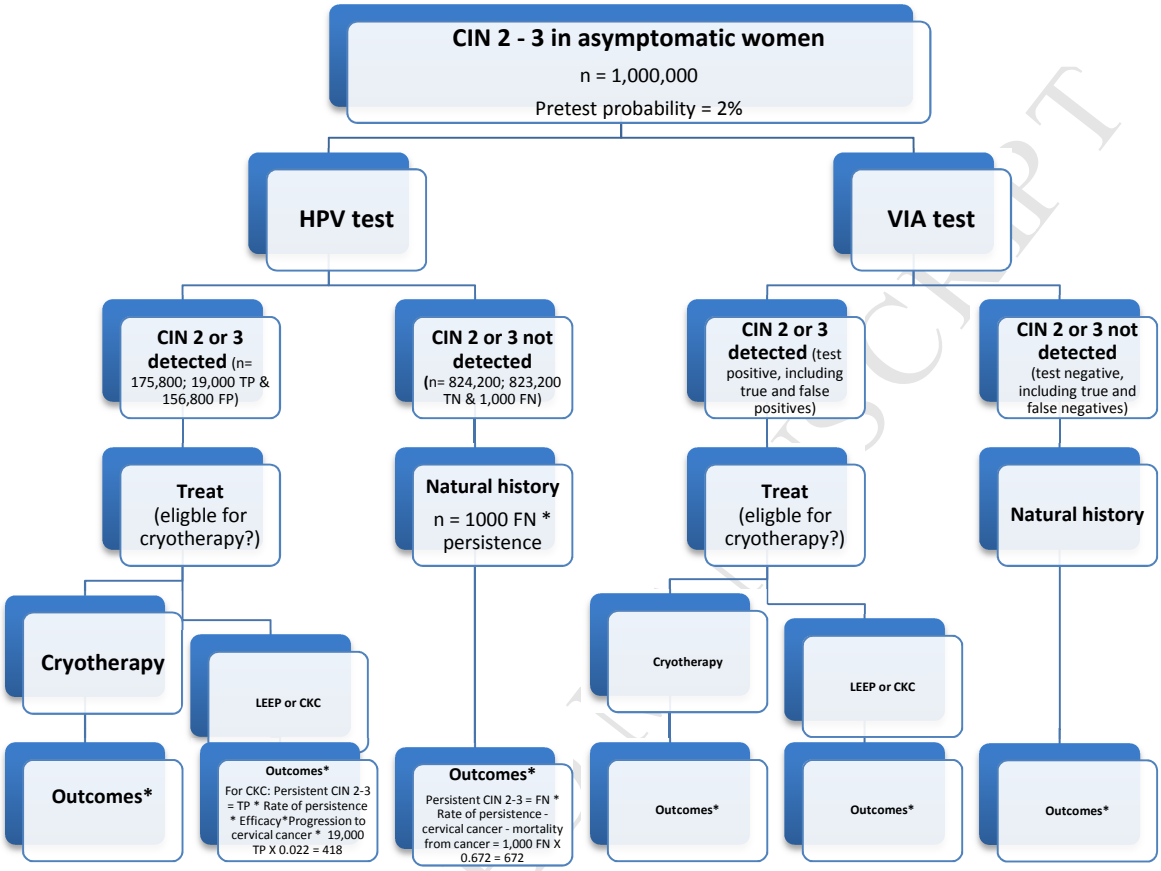
Figure 2. Generic analytic framework for a test from (2) from the USPTF – screening here refers to the application of a test (reprinted with permission)



Legend: numbers refer to key questions as follows: (1) Is there direct evidence that screening reduces morbidity and/or mortality? (2) What is the prevalence of disease in the target group? Can a high-risk group be reliably identified? (3) Can the screening test accurately detect the target condition? (a) What are the sensitivity and specificity of the test? (b) Is there significant variation between examiners in how the test is performed? (c) In actual screening programs, how much earlier are patients identified and treated? (4) Does treatment reduce the incidence of the intermediate outcome? (a) Does treatment work under ideal, clinical trial conditions? (b) How do the efficacy and effectiveness of treatments compare in community settings? (5) Does treatment improve health outcomes for people diagnosed clinically? (a) How similar are people diagnosed clinically to those diagnosed by screening? (b) Are there reasons to expect people diagnosed by screening to have even better health outcomes than those diagnosed clinically? (6) Is the intermediate outcome reliably associated with reduced morbidity and/or mortality? (7) Does screening result in adverse effects? (a) Is the test acceptable to patients? (b) What are the potential harms, and how often do they occur? (8) Does treatment result in adverse effects?

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Figure 3. Analytical framework for alternative pathways to screen and treat for cervical cancer. See also: https://dbep.gradepro.org/profile/50952068-76AE-516A-9CA3-5DEF08A61624



HPV = human papilloma virus
VIA = visual inspection with acetic acid
Test + = True and false positive tests (not known when test is performed)
Test - = True and false negatives (not known when test is performed)
CKC = Cold knife conization
Leep = Loop electrosurgical excision procedure
Cryo = cryotherapy
* Mortality from cervical cancer (including those detected with recurrence or unsuccessful treatment), incidence of cervical cancer (including those detected with recurrence or unsuccessful treatment), CIN 2-3, undetected CIN 2-3), cancers detected at treatment and complications of both necessary and unnecessary (false positives) treatment including major bleeding, premature deliver and infertility.

Examples of outcomes and analysis (time frame 30 years) – simplified modeling:

CIN 2-3 in HPV/CKC (both from recurrence and lack of detection)
Total # of women with outcome in HPV +/-CKC: 1090 resulting from a 2.2% persistence rate after CKC (19,000 TP X 0.022 = 418) and 67% persistence (1,000 FN X 0.672 as a result of 70% natural persistence rate minus cervical cancer and mortality from cervical cancer = 672)

Cervical cancer

CIN 2-3 = 20,000 (2% of 1,000,000) resulting from 19,000 TP and 1,000 FN
TP=19,000
FN=1,000
19,000 TP X 0.00055 (based on a systematic review of treatment for CKC treatment and progression to cervical cancer) = 11
1,000 FN X 0.0175 (from 70% natural persistence rate minus cervical cancer and mortality from cervical cancer and annual rate of progression to cervical cancer)
= 17
Total: 11+17=28

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Figure 4. Linking test accuracy to patient important outcomes

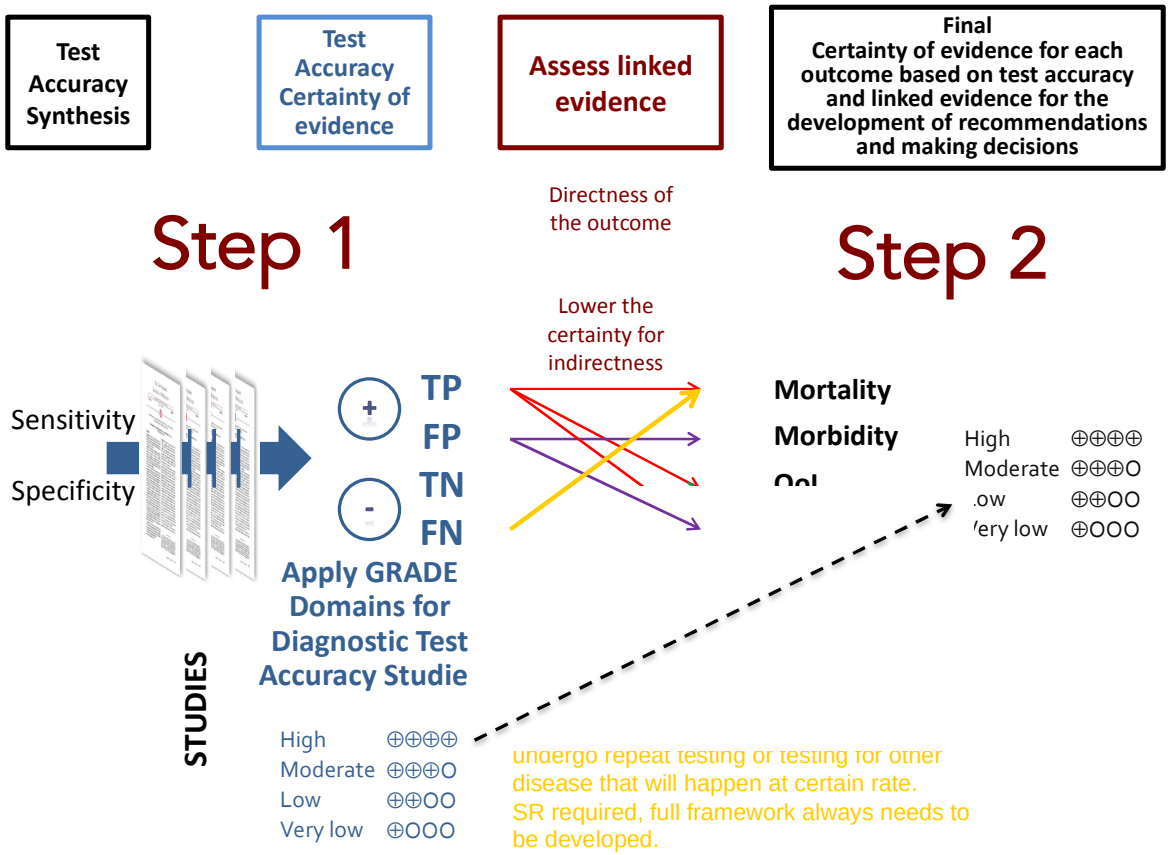
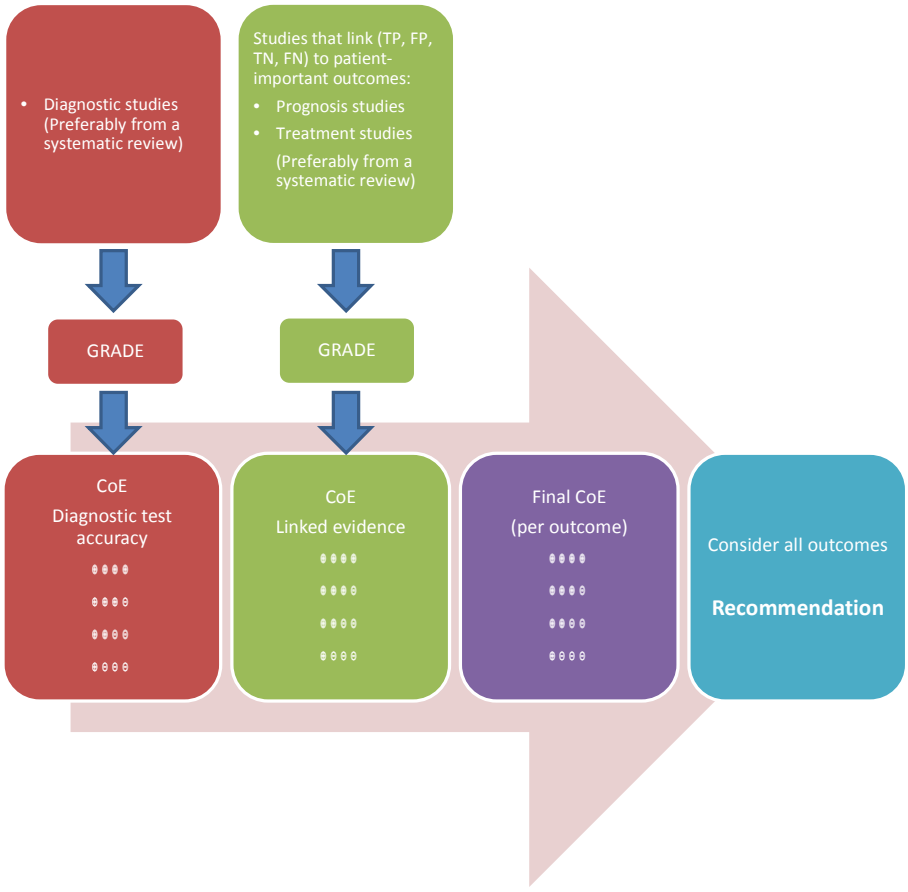


Figure 5. Linked evidence ratings that influence the overall certainty of the evidence.



CoE = certainty of evidence

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Table 1. Layer 1 SoF Table HPV compared to VIA for detection of cervical intraepithelial neoplasia in women at risk for cervical cancer – see https://dbep.gradepro.org/profile/50952068-76AE-516A-9CA3-5DEF08A61624 for GRADEpro interactive Summary of Findings Table

Patients or population: Women at risk of cervical cancer
Settings: screening clinics across the world
New Test: HPV **Cut-off value:** –
Comparison Test: VIA **Cut-off value:** –
Purpose: Screen and treat
Role: Replacement
Reference Test: conization and biopsy

Number of Participants (Studies)	8921 (5)	**Pooled Sensitivity HPV**	95% (95% CI: 84 to 98)	**Pooled Sensitivity VIA**	69% (95% CI: 54 to 81)
		Pooled Specificity HPV	84% (95% CI: 72 to 91)	**Pooled Specificity VIA**	87% (95% CI: 79 to 92)

Test Result	Number of results in 1000 per 1,000,000 women tested		Quality of the Evidence (GRADE)
	Baseline risk 2%[1]		
	HPV	VIA	
True positives (TP)	19 (17 to 20)	14 (11 to 16)	⊕⊕⊕⊕ **high**
TP absolute difference	**5 more**		
False negatives (FN)	1 (0 to 3)	6 (4 to 9)	
FN absolute difference	**5 less**		
True negatives (TN)	823 (706 to 892)	853 (774 to 902)	⊕⊕⊕⊖ **moderate**[2,3] due to inconsistency
TN absolute difference	**30 less**		
False positives (FP)	157 (88 to 274)	127 (78 to 206)	
FP absolute difference	**30 more**		

Reference: (15).

Footnotes:

[1] Prevalence of 2% was assumed to be the average prevalence in a representative population, numbers are rounded and expressed as women per million to reflect the population estimates and help with understanding the modeling in table 2
[2] Estimates of HPV and VIA sensitivity and specificity were variable despite similar cut-off values; inconsistency could not be explained by quality of studies. This was a borderline judgment. We downgraded TN and FP. This decision is considered in the context of other factors, in particular, imprecision.
[3] Wide CI for TN and FP that may lead to different decisions depending on which of the confidence limits is assumed.

Table 2. Layer 3 Summary of Findings Table describing population important outcomes (all certainty of evidence ratings are very low and therefore not included separately): see also: https://dbep.gradepro.org/profile/50952068-76AE-516A-9CA3-5DEF08A61624

Outcomes (quality of evidence very low for all of the outcomes)	Events in the screen-treat strategies for patient important outcomes (numbers presented per 1,000,000 women)						
	HPV +/- CKC	HPV +/- LEEP	HPV +/- Cryo	VIA +/- CKC	VIA +/- LEEP	VIA +/- Cryo	NO screen
Mortality from cervical cancer	20	30	30	81	88	88	250
Cervical Cancer Incidence	28	43	43	112	124	124	350
CIN2-3 recurrence	1088	1677	1677	4328	4762	4762	13400
Undetected CIN2-3 (FN)	1000			6000			N/A
Major bleeding	1511	397	60	1210	318	48	0
Premature delivery	712	575	610	670	560	588	500
Major infections	156	225	24	125	180	19	0
Minor infections	1649	1061	1139	1321	850	913	0
Unnecessarily treated (FP)	157000			127000			N/A
Cancer found at one time screening	2454			3168			N/A

The overall CoE for each of these outcomes is very low ⨁⊖⊖⊖. Our lack of confidence in these effect estimates stems mainly from very low quality evidence for treatment effects and natural progression/history data. Outcomes are not exclusive of each other but listed even if occurring twice as part of another (cervical cancer and mortality) to facilitate decision-making. Lighter cells (green) indicate more favorable estimates compared to darker (red) cells related to the option.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

We assume no mortality from cervical cancer in true negative (TN) and false positive (FP). To calculate the mortality from cervical cancer, we assumed 250 deaths per 350 women with cervical cancer. These numbers are based on Eastern Africa age standardized rates of cervical cancer and mortality provided by WHO at http://globocan.iarc.fr/ - Accessed 30 October 2012)

We assume no cervical cancer in TN or FP. To calculate cervical cancer incidence in women with persistent CIN 2/3, we assumed 350 cervical cancers per 14000 women who have persistent CIN 2/3 (i.e. FN). This incidence is based on Eastern Africa age standardized rate of cervical cancer of 350 cervical cancers per 1000000 women, of whom 2% have CIN 2/3 (20000 women with CIN 2/3, and a subsequent 30% regression for a total of 14000 with persistent CIN 2/3). This data is available from WHO at http://globocan.iarc.fr/ - Accessed 30 October 2012)

We assume no CIN2/3 in TN and FP. Our calculations in the model are based on 70% natural persistence of CIN 2/3 with no treatment (30% regression) in FN. The incidence of cervical cancer and mortality are also subtracted from the CIN 2/3 in FN (see above for calculations). TP are treated and recurrence rates of CIN 2/3 are 5.3% in cryotherapy and LEEP, and 2.2% in CKC.

We assumed major bleed would be 0 in TN and FN as they were not treated. We assumed that a proportion of 0.000339 of the population treated with cryotherapy, 0.002257 with LEEP, and 0.008585 with CKC, based on pooled proportions in observational studies with no independent controls, will have major bleeding.

We assumed 5% population risk of premature delivery in 1% women who become pregnant. Based on pooled meta-analysis of controlled observational studies, 0.001125 of the population treated with cryotherapy, 0.000925 with LEEP, and 0.001705 of the population treated with CKC will have premature delivery.

We did not identify any data about the risk of infertility after treatment for CIN2+.

We assumed major infection would be 0 in TN and FN as they were not treated. Based on pooled proportions from studies with no independent control 0.000135 of the population treated with cryotherapy 0.001279with LEEP, and 0.000888 with CKC will have major infection.

We assumed minor infection would be 0 in TN and FN as they were not treated. Based on pooled proportions from studies with no independent control, 0.006473 of the population treated with cryotherapy, 0.006027 with LEEP, and 0.009368 with CKC will have minor infection.

Cancers detected at screening was calculated as the average number of cancers detected when screening across all diagnostic studies which contributed to the DTA data for each comparison.

No screen numbers were calculated using the same assumptions above for FN, with the exception of premature delivery which was baseline risk in the population.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Table 3. GRADE SoF tables for test accuracy and linked evidence

GRADE SoF Layers	Information included	End users
One	Summarize the results of test accuracy reviews and certainty of those estimates.	Layer 1 is usually used by systematic reviewer authors.
Two	Summarize the results of test accuracy reviews (layer 1) in addition to estimates of the direct adverse consequences, inconclusive results and certainty of those estimates.	Layer 2 includes assumptions about consequences for patient important outcomes
Three	Summarize the modeled effects of performing the	Layer 3 should be used by guideline developers or

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

	test on important outcomes based on the best available research evidence.	other decision-makers. Systematic reviewers of test accuracy would require additional information about the linked evidence and modeling.

C:\pdfconversion\WORK\authordoc\Article 22 JCE series diagnosis 2 20181113 clean.docx

Highlights

GRADE has developed and applied a comprehensive framework to evaluate the certainty of the evidence when direct evidence of the effect of testing on outcomes is not available and linked evidence that connects test accuracy to downstream consequences is required for decision-making. Ideally, this linked evidence comes from systematic reviews is that informs analytical frameworks for questions related to tests.

What this adds to what is known?

Application of GRADE's approach requires guideline developers to rate their certainty in each element of the linked evidence that is required for decision-making.

What are the implications, what should change now?

Further research should address ways to arrive at, for each critical or important outcome and across outcomes, ratings of the certainty derived from the linked sources of evidence. This will often require formal modelling and assessing the certainty in the models.

Disclosure Statement

The authors are members of the GRADE Working Group. JAS has received research grants from Takeda and Savient and consultant fees from Savient, Takeda, Regeneron, Merz, Iroko, Bioiberica, Crealta/Horizon and Allergan pharmaceuticals, WebMD, UBM LLC and the American College of Rheumatology. JAS serves as the principal investigator for an investigator-initiated study funded by Horizon pharmaceuticals through a grant to DINORA, Inc., a 501 (c)(3) entity. JAS is a member of the executive of OMERACT, an organization that develops outcome measures in rheumatology and receives arms-length funding from 36 companies; a member of the American College of Rheumatology's (ACR) Annual Meeting Planning Committee (AMPC); Chair of the ACR Meet-the-Professor, Workshop and Study Group Subcommittee; and a member of the Veterans Affairs Rheumatology Field Advisory Committee. JAS is the editor and the Director of the UAB Cochrane Musculoskeletal Group Satellite Center on Network Meta-analysis. The other authors declared no financial conflict of interest.