ELSEVIER

Journal of Clinical Epidemiology ■ (2019) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper

Monica Hultcrantz[a,*], Reem A. Mustafa[b,c], Mariska M.G. Leeflang[d], Valéry Lavergne[e,f], Kelly Estrada-Orozco[g,h], Mohammed T. Ansari[i], Ariel Izcovich[j], Jasvinder Singh[k,l,m], Lee Yee Chong[n], Anne Rutjes[o], Karen Steingart[p], Airton Stein[q,r], Nigar Sekercioglu[s,c], Ingrid Arevalo-Rodriguez[t], Rebecca L. Morgan[c], Gordon Guyatt[c], Patrick Bossuyt[d], Miranda W. Langendam[d], Holger J. Schünemann[c,u]

[a]*Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), S:t Eriksgatan 117, SE-102 33, Stockholm, Sweden*
[b]*Division of Nephrology and Hypertension, Department of Medicine, University of Kansas Medical Center, 3901 Rainbow Blvd, MS3002, Kansas City, KS 66160, USA*
[c]*Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[d]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Public Health research institute, Amsterdam University Medical Centers, Meibergdreef 9, Amsterdam, The Netherlands*
[e]*Department of Medicine, Sacré-Coeur Hospital, University of Montreal, Montreal, Canada*
[f]*Department of Clinical Affairs & Practice Guidelines, Infectious Disease Society of America, Arlington, VA, USA*
[g]*Clinical Research Institute, Faculty of Medicine, Universidad Nacional de Colombia, Bogotá, Colombia*
[h]*Research Unit, Faculty of Medicine, Fundación Universitaria Sanitas, Bogotá, Colombia*
[i]*School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, 600 Peter Morand Crescent, Ottawa, Ontario K1G 5Z3, Canada*
[j]*Internal Medicine Service, German Hospital, Pueyrredón 1640, Buenos Aires C1118AAT, Argentina*
[k]*Medicine Service, VA Medical Center, 510, 20th street South, FOT 805B, Birmingham, AL, USA*
[l]*Department of Medicine at the School of Medicine, University of Alabama at Birmingham (UAB), 1720 Second Ave South, Birmingham, AL 35294-0022, USA*
[m]*Department of Epidemiology at the UAB School of Public Health, 1665 University Blvd., Ryals Public Health Building, Room 220, Birmingham, AL 35294-0022, USA*
[n]*Ateimed Consulting Ltd., 3rd Floor 166 College Road, Harrow, Middlesex HA1 1BH, UK*
[o]*Institute of Social and Preventive Medicine (ISPM), University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland*
[p]*Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK*
[q]*Programa de Ciências da Saúde - Universidade Federal de Ciências da Saúde de Porto Alegre (Ufcspa) Rua Sarmento Leite, 245 - CEP 90050-170, Porto Alegre, Brazil*
[r]*Teaching and Research Unit - Grupo Hospitalar Conceição (GHC) Rua Francisco Trein, 596 - CEP - 91.350-200, Porto Alegre, Brazil*
[s]*Division of Nephrology, Department of Medicine, University of Toronto, 585 University Avenue, Toronto, Ontario M5G 2N2, Canada*
[t]*Clinical Biostatistics Unit, Hospital Universitario Ramon y Cajal, CIBER Epidemiology and Public Health, Ctra. Colmenar Km. 9,100, 28034 Madrid, Spain*
[u]*Department of Medicine, Michael G DeGroote Cochrane Canada Centre and GRADE centre, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*

Accepted 14 May 2019; Published online xxxx

Abstract

Objective: The objective of the study was to clarify how the Grading of Recommendations Assessment, Development and Evaluation (GRADE) concept of certainty of evidence applies to certainty ratings of test accuracy.

Study Design and Setting: After initial brainstorming with GRADE Working Group members, we iteratively refined and clarified the approaches for defining ranges when assessing the certainty of evidence for test accuracy within a systematic review, health technology assessment, or guideline.

Conflicts of interest: All authors are members of the GRADE Working Group. M.H., R.A.M., M.M.G.L., M.W.L., H.J.S., V.L., J.S., M.T.A., K.E-O., A.I., L.Y.C., A.R., A.S., N.S., I.A.R., R.L.M., G.G., and P.B. have nothing to disclose. K.S. reports to have received financial support for the preparation of systematic reviews and educational materials, consultancy fees from FIND (for the preparation of systematic reviews), honoraria, and travel support to attend World Health Organization guideline meetings.

* Corresponding author. Tel.: +46 (0)8-412 32 73; fax: +46 (0)8-411 32 60.
E-mail address: monica.hultcrantz@sbu.se (M. Hultcrantz).

https://doi.org/10.1016/j.jclinepi.2019.05.002

Results: Ranges can be defined both for single test accuracy and for comparative accuracy of multiple tests. For systematic reviews and health technology assessments, approaches for defining ranges include some that do not require value judgments regarding downstream health outcomes. Key challenges arise in the context of a guideline that requires ranges for sensitivity and specificity that are set considering possible effects on all critical outcomes. We illustrate possible approaches and provide an example from a systematic review of a direct comparison between two test strategies.

Conclusions: This GRADE concept paper provides a framework for assessing, presenting, and making decisions based on the certainty of evidence for test accuracy. More empirical research is needed to support future GRADE guidance on how to best operationalize the candidate approaches.

Keywords: Certainty of evidence; Test accuracy; GRADE; Guidelines; Systematic reviews; Health technology assessments

1. Introduction

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) concept of certainty of evidence (also called quality of evidence) represents our confidence that the true effect lies above or below a threshold, or in a specified range [1]. To assess the certainty of evidence for an individual outcome, authors of systematic reviews, health technology assessments (HTAs), or guidelines need to specify the thresholds or ranges they are using and the associated rationale. Several approaches exist for setting thresholds and ranges. For recommendations in clinical practice and public health guidelines, GRADE has suggested setting a threshold based on consideration of all critical outcomes. For systematic review authors, we have illustrated three different approaches: expressing certainty in the range set by the 95% confidence interval (CI), certainty in the direction of effect, or certainty in a particular magnitude of effect, for example, small, medium, or large.

Although GRADE has illustrated the concept of certainty using effects of treatment interventions, the guidance to specify ranges or thresholds also applies to questions of diagnostic tests. When diagnostic intervention studies comparing alternative diagnostic test strategies with direct assessment of patient-important outcomes are available (such as RCTs addressing the impact on survival after a screening strategy), the approaches for setting thresholds or ranges previously presented apply [1]. In this paper, we will explore the concepts when there are no such studies.

If no studies have directly compared the effects of alternative test strategies on downstream health outcomes, modeling the impact of diagnostic accuracy on the health outcomes could inform management decisions [2,3]. For example, false-negative (FN) and false-positive (FP) test results, by missing or delaying the diagnosis (FN) or through unnecessary treatment (FP), can adversely impact health outcomes [2,3]. GRADE previously described that to evaluate impact, one may, through formal or informal modeling, link different types of evidence: diagnostic test accuracy estimates (e.g., sensitivity and specificity), direct effects of the test(s) (e.g., complications of an invasive test), natural course of the condition, treatment effectiveness, and the link between the test results and clinical management [2–4]. Arriving at an overall rating of certainty of evidence requires rating every component.

This article explores possible ways of setting thresholds or ranges for rating certainty in diagnostic test accuracy, and what this would mean in the context of systematic reviews, HTA, and health care recommendations. GRADE has described approaches for setting thresholds or ranges in terms of levels of contextualization [1]. Box 1 presents levels of contextualization for diagnostic accuracy, concepts that this paper will further illustrate. The discussion is consistent with previous guidance on rating certainty in diagnostic accuracy [2–4].

2. Definitions and scope

In our previous work clarifying the construct of certainty of evidence, we used the term *threshold* as a set border (e.g., a threshold at which the benefits start outweighing the harms) and the term *range* when using two borders (e.g., the upper and lower limits of a small effect). Although one could use the same terminology for the borders set in test accuracy, to avoid confusion with the thresholds used to dichotomize the test results for a particular test, throughout this paper, we will use range meaning *threshold or range*.

We use the term *test strategy* to denote a combination of tests (e.g., clinical test followed by magnetic resonance imaging), not to be confused with test-treatment strategy that also includes the treatment that is guided by the test result [2]. The test under consideration can have different roles within a test strategy: to replace an existing test, as triage test before an existing test, as an add-on to an existing test [5], or parallel to an existing test [6]. When evaluating diagnostic accuracy of a test, it is important to define the role of the test to address the accuracy of the full test strategy. The approaches for setting ranges suggested in this paper apply to all types of test strategies. We will present the approaches for comparisons between tests as well as for single tests, but our main focus will be on the comparative scenario, which we will further explain below.

What is new?

Key findings

- This Grading of Recommendations Assessment, Development and Evaluation (GRADE) concept paper shows that the choice of ranges is important when rating the certainty of evidence for test accuracy because it may affect the interpretation of the result and the degree of certainty presented.
- We present possible approaches for setting ranges for sensitivity and specificity for a single test and for a comparison of test options. The approaches are illustrated using an example of a direct comparison between two test strategies.

What this adds to what was known?

- The GRADE Working Group has previously clarified that the concept of certainty of evidence represents our confidence that the true effect lies above or below a threshold, or in a specified range. The frequent lack of direct evidence assessing the effect of medical tests on patient important outcomes highlights the need for a clarification of how these concepts apply to certainty ratings of test accuracy.

What is the implication and what should change now?

- When rating the certainty of evidence for test accuracy, it is important that systematic review authors, health technology assessors, and guideline developers are transparent with the ranges they are using, the rational for choosing them and with the value judgments made.
- More empirical data are needed before knowing which approaches, for defining ranges of accuracy, would be most useful for different purposes, and how to best operationalize them.

Box 1 Degree of contextualization when defining range

- Noncontextualized (primarily for systematic reviews and health technology assessments). The ranges used are independent of value judgments regarding, for example, the relative importance of false negatives vs. false positives.
- Partially contextualized (primarily for systematic reviews and health technology assessments). The ranges depend on some value judgment—for example, the importance of downstream health consequences of true and false positives and negatives. This approach to contextualization requires setting boundaries of ranges expressed in absolute terms for a given prevalence.
- Fully contextualized (primarily for guidelines and other decision making). The boundaries are set considering the range of possible effects on all critical outcomes, bearing in mind the decision(s) that need to be made and the associated values and preferences. This approach to contextualization requires setting boundaries of ranges expressed in absolute terms for a given prevalence.

When addressing the certainty of evidence for test accuracy, we are presenting and rating ranges for sensitivity and specificity. However, when interpreting a test result in clinical practice, multilevel likelihood ratios or multivariable approaches may be more useful.

We refer to noncontextualized certainty ratings if authors make choices of ranges without value judgments that do not involve modeling. The term fully contextualized refers to situations in which the entire health care question/context is considered when assessing the certainty of sensitivity and specificity, typically in the setting of a guideline [1]. Less-contextualized ratings are typically made in systematic reviews and HTA. We will continue to make distinctions between certainty ratings that are fully contextualized (considering all critical outcomes with their associated values within a particular decisional context), partly contextualized (including some value judgment regarding the importance of the individual outcome), and noncontextualized (without value judgments). Noncontextualized or partially contextualized approaches refer only to the chosen ranges and not to other decisions. For instance, authors of systematic reviews always need to consider the context of interest, for example, in their eligibility criteria (e.g., only including studies with a certain prevalence or setting), or when assessing indirectness.

Currently, authors use decision models of varying complexity to inform decisions regarding test strategies: ranging from back of the envelope estimations of the possible consequences to advanced models estimating all benefits and harms to the patients as well as the uncertainty associated with the parameters in the model [3]. We will exemplify the contextualized approaches using a simple model estimating the consequences of changes in the sensitivity and specificity of the test strategies. However, the concepts we present apply to any level of modeling, requiring only consideration of all critical direct and downstream outcomes.

3. Comparisons of test strategies

If the goal is to evaluate two test strategies, one can compare the accuracy of the two tests using a study design

in which one administers the tests in the same population comparing to the same reference standard (direct comparison) [7]. In many cases, however, primary research has only studied the accuracy of single tests against a reference standard in separate populations and separate studies. In these cases, the comparison between the relevant tests will be indirect, leading to additional challenges beyond the scope of this paper.

Table 1 shows possible approaches for setting ranges in sensitivity and specificity and illustrates what the certainty ratings represent for a direct comparison vs. a single test. We will start by presenting an overview of the suggested approaches and then continue with an example of applying the approaches to a direct comparison.

4. Noncontextualized ratings of test accuracy (typically for systematic reviews and health technology assessments)

We refer to the first two approaches presented in Table 1 as noncontextualized, meaning that the choice of the boundaries for the range of sensitivity and specificity does not involve value judgments (Box 1). That is, the importance of the number of FNs or FPs does not bear on the ranges chosen, and the downstream consequences of the test results have no influence on the certainty ratings of sensitivity and specificity. Analysts use these approaches when they wish to assess the certainty of the test accuracy without further interpreting the results or providing advice.

Table 1. Possible ways of setting ranges for sensitivity and specificity and what the certainty expressed will represent for a comparison between tests vs. single test

			What the certainty rating represents	
Degree of contextualization	**Range**	**How it is set**	**For a comparison between tests**	**For a single test**
Noncontextualized (primarily for systematic reviews and health technology assessment)	Range: 95% Confidence Interval	Using existing limits of the 95% CIs, which implies precision is not routinely part of the rating	Certainty that the true difference in accuracy lies within the confidence region of the tests compared or the true difference in sensitivity and specificity lies within their respective confidence intervals	Certainty that the true sensitivity and specificity lies within their respective confidence intervals
	Difference ≠ 0	Using the threshold of null effect	Certainty that the sensitivity or specificity of one test strategy differs from that of another	Not applicable
Partially contextualized (primarily for systematic reviews and health technology assessment)	Specified magnitude (set in natural frequencies for a given prevalence)	For example, a small difference in sensitivity or specificity can be defined as a difference small enough that one might consider not using the test if adverse effects or costs are appreciable	Certainty in a specified magnitude of difference between the sensitivity or specificity of two tests (e.g., no or trivial, small, medium, or large difference)	Certainty in a specified magnitude of sensitivity or specificity (e.g., low, moderate, or high accuracy)[a]
Fully contextualized (primarily for guidelines)	Range determined with considerations of all critical direct and downstream health outcomes or all desirable and undesirable consequences (set in natural frequencies for a given prevalence)	Considering the range of possible effects on all critical health outcomes and consequences [3], bearing in mind the decision(s) that need to be made, and the associated values and preferences	For each outcome (in this case sensitivity and specificity), ratings represent our confidence that the overall balance between net benefit and net harm will not differ from one end of the certainty range[b] to the other.	For each outcome (in this case sensitivity and specificity), ratings represent our confidence that the overall balance between net benefit and net harm will not differ from one end of the certainty range[b] to the other.

[a] This will have to be specific to the test, condition, and setting. Sensitivity of 97% may be considered extremely accurate for some conditions/test/setting, whereas it may be considered low accuracy for a different scenario. This decision will be informed by the patient and population consequences based on the test results.

[b] By certainty range, we mean the range in which we anticipate that the true sensitivity or specificity may lie, after considering not only precision but also risk of bias, inconsistency, indirectness, and publication bias [1,8].

4.1. Using the ranges of the confidence intervals

The first approach assesses how certain we are that the true sensitivity and specificity lies within the observed CIs. Using this approach, one omits the rating of imprecision, that is, one could have high certainty that the true sensitivity or specificity lies within the range set by the CI regardless of whether this range is wide or narrow. The ranges can be presented for sensitivity and specificity, or for the number of FPs and FNs, given a particular pretest probability. In comparing two tests, one will rate the certainty of the difference in sensitivity and specificity or FPs and FNs between the tests under consideration. This approach could potentially mean that we express high certainty in very imprecise results.

4.2. Using the direction of effect

The second approach assesses our certainty regarding whether a difference exists between the accuracy of two test strategies. In other words, how certain are we that test A has a higher/lower sensitivity or specificity than test B? In some cases, one would want to address the certainty that the true difference in test accuracy lies close to no difference. This requires a decision regarding what difference would be trivial and thus requires a partly contextualized judgment that we describe in the following.

5. Partly contextualized ratings of certainty: ranges of magnitude of accuracy (typically for systematic reviews and health technology assessments)

The third option described in Table 1 is to rate our certainty in a specific accuracy. When applying this approach to a comparison between two tests, one could specify categories of no or trivial, small, moderate, or large difference in accuracy. Similarly, when evaluating the accuracy of a single test in comparison to the reference standard, one could specify trivial, low, moderate, or high accuracy. This approach requires setting boundaries of ranges expressed in absolute terms for a given prevalence—boundaries that likely will depend on the value placed on the direct effects (i.e., burdens/adverse effects) of the test as well as the downstream health consequences of the true and false positives and negatives.

For example, consider a situation in which the downstream health consequences of a management decision are serious, such as recurrence of disease. In such situations, ranges of FPs and FNs will have a lower value than if the downstream consequences are less serious such as minor adverse events or length of hospital stay.

6. Fully contextualized ratings (typically for guidelines) of test accuracy

When we make fully contextualized ratings, we are simultaneously weighing the benefits and harms of every critical or important health outcome or even all desirable and undesirable consequences [1] (Box 1). In the absence of studies comparing the health consequences of tests, one would ideally use a fully developed model for assessing the effects of the test strategies on patient important outcomes. If such models can generate estimated effects with CIs, one can make fully contextualized ratings of the patient important outcomes in the same way as we have previously described [1]. The accuracy data would in this case be one of several pieces of data feeding into the model.

Currently, guideline panels seldom have access to advanced models. As a result, they will inevitably focus on diagnostic accuracy [9−11]. Here, we discuss how one can, in these cases, make fully contextualized ratings of sensitivity and specificity, that is, address whether one would make a different decision at either end of the certainty ranges. One can then use models or explicit considerations to decide what sensitivity and specificity one would require to recommend a particular test. That is, what levels of sensitivity and specificity would be required to ensure that the desirable health effects will outweigh the undesirable. In some cases, it is also possible to set ranges for sensitivity and specificity by inferring decision thresholds from other recommendations and decisions about testing [12].

When all else is judged exactly equal between the two test strategies (e.g., side effects, invasiveness, resources considerations, timing of test, location of test in the care pathway, feasibility, availability), the fully contextualized range would be the same as the noncontextual no-effect range. Although it is unlikely to occur, one could then base a decision solely on knowledge of whether the test accuracy increases or decreases with one test-strategy compared with another [9,13,14].

Fully contextualized ranges are often decided on through discussions in guideline panels, based on what is known about the direct and downstream health outcomes of the test strategies. In some cases, panels conduct formal surveys of their members to establish test and treatment thresholds [15]. In some situations, considering downstream health outcomes can be sufficient and no formal modeling is needed—for example, if it is obvious that the health consequences of using the test would be negative.

In this paper we will illustrate how simple models of health outcomes can inform the choice of fully contextualized ranges for sensitivity and specificity. If the values are very uncertain, or the results will be used in several different contexts, one can provide several certainty ratings, each for a specific set of values.

7. Applying ranges to direct comparisons of accuracy between test strategies

To make decisions about tests, direct comparisons of the relevant test strategies are ideal. We will show what the

approaches for setting ranges would mean in such a setting using the direct comparison of accuracy between two tests for cervical cancer screening, the human papillomavirus (HPV) test (HPV DNA-PCR testing) and unaided visual inspection of the cervix with acetic acid (VIA) [16].

Cervical intraepithelial neoplasia (CIN) is a premalignant lesion diagnosed by histology, in three stages: CIN 1, CIN 2, and CIN 3. If left untreated, CIN 2 or 3 (CIN 2-3) can progress to cervical cancer. HPV causes virtually all cancer of the cervix and is the most common sexually transmitted disease [17]. The setting for this example is a screen-treat strategy in low- and middle-income countries, in which treatment is provided to all with a positive screening test.

We used this example in prior GRADE articles [2,3,5]. It is based on a systematic review of five studies assessing the accuracy of HPV and VIA against a common reference standard (a combination colposcopy with or without biopsy and in some instances clinical follow-up as well) [16]. For the HPV test, the pooled sensitivity was 95% (95% CI: 84 to 98) and the pooled specificity 84% (95% CI: 72 to 91), and for VIA, the pooled sensitivity was 69% (95% CI: 54−81) and the pooled specificity 87% (95% CI: 79 to 92). The diagnostic sensitivity is 26% points higher with HPV compared with VIA (95% CI: 11% to 41% higher), whereas the specificity is 3% lower (95% CI: 15% lower to 8% higher) (Method in Appendix 1). At the estimated prevalence in the general population of 2%, based on WHO data [17], if 1,000 women are screened with the HPV test instead of VIA, five more true positives will be found (2 to 8 more), although there will be 34 more FPs (147 more to 78 fewer), who would receive treatment.

No serious concerns regarding risk of bias, indirectness, or publication bias for this comparison of test accuracy were identified. Whether there are serious problems with inconsistency (the estimated differences in sensitivity with HPV vs. VIA in the included five studies ranged from an increase of 1% to an increase in 56%, and the estimated difference in specificity ranged from a decrease of 22% to an increase in 9%) and imprecision will depend on the ranges used; those judgments are described in the following.

7.1. Noncontextualized approaches (primarily for systematic reviews or health technology assessments)

7.1.1. Using the ranges of the confidence intervals

The first noncontextualized approach listed in Table 1 is to assess our certainty in the ranges defined by the 95% CIs. In this case, we would be rating how certain we are that the sensitivity of the HPV test is 11% to 41% higher than VIA and the specificity is somewhere between 15% lower and 8% higher. Because the estimated difference of the test accuracy results in individual studies are outside of these ranges, we might rate down for inconsistency in both sensitivity and specificity. With this approach, we do not judge the width of the intervals, that is, precision is omitted from the ratings, and because no serious concerns were identified for the other domains, we would end up with moderate ratings of certainty for the ranges set by the 95% CIs (Table 2). Different target audiences can use these ranges with certainty ratings for their particular goals, for example, as input into a model for estimating downstream consequences.

7.1.2. Using the direction of effect

The second noncontextualized approach for defining ranges is to use the boundary of no difference in sensitivity or specificity. When doing so, we are addressing our certainty in the direction (increase or decrease) of sensitivity and specificity, neglecting the magnitude of the difference. In this case, we would be rating how certain we are that by using the HPV test rather than VIA, we would increase the sensitivity and decrease the specificity. Because the entire CI for the difference in sensitivity lies on one side of no effect, as well as the estimated differences in all the included studies, we would not rate down for imprecision or inconsistency and we would have high certainty that the HPV test indeed increases the sensitivity for detecting CIN 2-3. On the other hand, there is a serious problem with imprecision for specificity because the CI reaches from a decrease of 15% to an increase of 8%. Furthermore, individual studies have estimated differences in specificity between an increase in 9% and a decrease in 22%, and we would therefore rate down for both imprecision and inconsistency (Table 2).

7.1.3. Partly contextualized ratings of certainty: ranges of magnitude of accuracy (primarily for systematic reviews and health technology assessments)

Using this approach, one would define ranges for a trivial, small, moderate, or large difference in sensitivity and specificity. Because these judgments are based on the downstream health consequences, reviewers must address these clearly in the beginning of the review process. In our example, a simple model was used based on five outcomes: cervical cancer, cervical cancer-related mortality, major bleeds, premature delivery, and major infections (Fig. 1; detailed explanation in Appendix 1). Cervical cancer and cervical cancer-related mortality due to FN test results could be reduced using a test with a higher sensitivity. Major bleeds, premature delivery, and major infections due to FP test results can be reduced using a test with a higher specificity.

The model provides approximations regarding how differences in sensitivity and specificity will affect the outcomes of interest. At a prevalence of 2% [17], increasing sensitivity by 1% would result in approximately two fewer cervical cancer-related deaths and three fewer cases of cervical cancer per million women screened. A 1% increase in specificity will result in approximately three fewer major bleeds, six fewer premature births, and 1 less major

Table 2. Examples of certainty ratings for the difference in sensitivity and specificity between HPV and VIA

Approaches	Examples of set ranges	Certainty
Range: 95% confidence interval	Sensitivity: The 95% CI, in this case, an increase by 11–41% (at a pretest probability of 2%, 2–8 more true positives per 1,000 women screened.) Specificity: The 95% CI, in this case, a decrease by 15% to increase by 8% (at a pretest probability of 2%, 147 more to 78 fewer false positives per 1,000 women screened)	We have moderate certainty that the true increase in sensitivity is between 11% and 41% (rating down for inconsistency) We have moderate certainty that the true difference in specificity is between a 15% decrease and an 8% increase (rating down for inconsistency)
rsens $\neq$ 1, rspec $\neq$ 1	Direction of effect	We have high certainty that the sensitivity of HPV testing is higher than VIA for detecting CIN 2-3 We have low certainty that the specificity of HPV testing is lower than VIA for detecting CIN 2-3 (rating down for imprecision and inconsistency)
Specified magnitude of difference in sensitivity and specificity	The set range for a large effect on sensitivity was a difference in more than four true positives per 1,000 screened (corresponds to mortality of more than 50 and cervical cancer cases of more than 60, per million screened) The set range for a no or trivial difference in specificity was a difference in up to 200 false positives per 1,000 screened (corresponds to a difference of up to approximately 33 major bleeds, 120 premature births, and 13 major infections per million women screened)	We have low certainty that HPV has a large increase in sensitivity compared with VIA (rating down for imprecision and inconsistency) We have low certainty that there is no or trivial difference in specificity between HPV and VIA (rating down for imprecision and inconsistency)
Range determined with considerations of all critical direct and downstream health outcomes or all desirable and undesirable consequences	Thresholds based on the values we place on mortality and cervical cancer vs. major bleeds, premature delivery, and major infections.	Considering downstream health outcomes, we have low certainty in the sensitivity outcome, that is, this outcome may not shift the overall balance between net benefit and net harm (rating down for imprecision and inconsistency). Considering downstream health outcomes, we have low certainty in the specificity outcome, that is, this outcome may not shift the overall balance between net benefit and net harm (rating down for imprecision and inconsistency).

Abbreviations: CI, confidence interval; CIN, Cervical intraepithelial neoplasia; HPV, human papillomavirus; VIA, visual inspection of the cervix with acetic acid.

The partially and fully contextualized ranges are set considering a prevalence of 2%.

infection per million women screened. One can use this information to guide the choice of ranges for no or trivial, small, moderate, or large difference in sensitivity or specificity. As previously noted, the choices of ranges will likely differ depending on the value placed on the outcomes. In contrast to a range set directly on a patient-important outcome, the ranges for sensitivity and specificity can be affected by several downstream health outcomes. This is illustrated in Table 3, in which examples of ranges for the difference in sensitivity and specificity for HPV vs. VIA are presented.

For this example, the point estimate for sensitivity was within the presented range of a large increase. As the CI crosses the border of a moderate increase, and individual studies have estimated differences that can be considered trivial or small, the certainty rating is low because of imprecision and inconsistency. For specificity, the point estimate is within the range defined as a no or trivial difference. Because the CI crosses the border to a small decrease, one would rate down for imprecision. In addition, one of the included studies has an estimated decrease of 22% in specificity (considered a medium-large difference), which

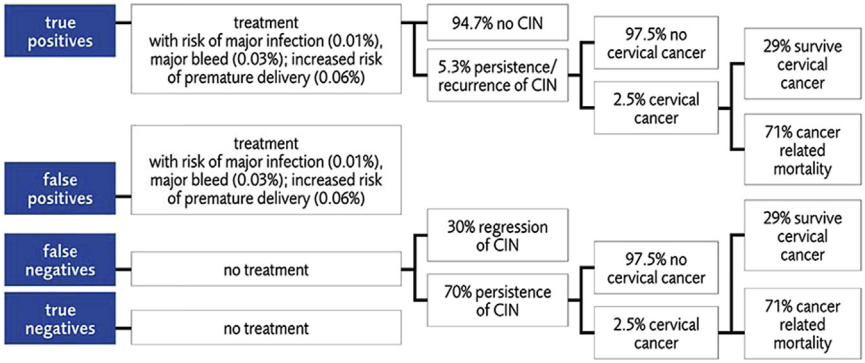

Fig. 1. Estimated consequences of the four possible test results for CIN. The setting for this example is a screen-treat strategy in low- and middle-income countries, in which treatment is provided to all with a positive screening test.

might warrant rating down for inconsistency, in which case, the certainty rating would be low (Table 2).

7.1.4. Fully contextualized ratings (primarily for guidelines or other decisions) of test accuracy

When making fully contextualized ratings of the difference in sensitivity and specificity, we start by considering the downstream health outcomes (Fig. 1). Moreover, just as for treatment interventions, we will have to specify values for all critical health outcomes. The values should be those of the patients, and the process for obtaining them can include a systematic review of the relevant literature, the experience of the topic experts in conducting shared decision-making, consultation with patients and patient groups, and conduct of targeted surveys [18–20].

In the present example, the guideline panel might infer that women eligible for screening would value major infections and major bleeds equally, premature delivery twice as high, and would place an appreciably greater value on cervical cancer and cervical cancer-related mortality, say seven and 20 times higher, respectively. Such an inference may be informed by, for example, reported utility estimates from similar clinical contexts [21–23].

The question will be how much harm we are willing to accept, given a certain benefit, or the other way around. For this particular example, the guideline panel will consider how certain they are that the increase in sensitivity is high enough to outweigh the potential decrease in specificity. At a prevalence of 2%, the estimated effect of increasing sensitivity with 1% is 2.5 fewer cervical cancer-related deaths

Table 3. Example of ranges set for different magnitudes of difference in sensitivity and specificity for HPV vs. VIA[a]

Sensitivity	Specificity
No or trivial difference in sensitivity: 0–4% Difference in 0–1 TP found per 1,000 screened (corresponds to a difference in mortality of up to 10 and cervical cancer cases of up to 12, per million women screened)	No or trivial difference in specificity: 0–10% Difference in 0–100 FP per 1,000 screened (corresponds to a difference of up to approximately 33 major bleeds, 60 premature births, and 13 major infections per million women screened)
Small difference in sensitivity: 4–10% Difference in 1–2 TP per 1,000 screened (corresponds to a difference in mortality of around 10–25 and cervical cancer cases of around 12–30, per million women screened)	Small difference in specificity: 10–20% Difference in 100–200 FP per 1,000 screened (corresponds to a difference of approximately 33–66 major bleeds, 60–120 premature births, and 13–26 major infections per million women screened)
Moderate difference in sensitivity: 10–20% Difference in 2–4 TP per 1,000 screened (corresponds to a difference in mortality of around 25–50 and cervical cancer cases of around 30–60, per million women screened)	Moderate difference in specificity: 20–30% Difference in 200–300 FP per 1,000 screened (corresponds to a difference of approximately 66–100 major bleeds, 120–180 premature births, and 26–39 major infections per million women screened)
Large difference in sensitivity: more than 20% More than 4 TP per 1,000 screened (corresponds to mortality of more than 50 and cervical cancer cases of more than 60, per million screened)	Large difference in specificity: more than 30% More than 300 FP per 1,000 screened (corresponds to a difference of approximately 100 or more major bleeds, 180 or more premature births, 39 or more major infections per million women screened)

Abbreviations: FP, false positives; HPV, human papillomavirus; TP, true positives; VIA, visual inspection of the cervix with acetic acid.

The boundaries of the ranges represent a hypothetical group consensus based on the importance placed on cervical cancer and cervical cancer-related mortality (for sensitivity), and major bleeds, premature births, and major infections (for specificity).

[a] The values for sensitivity and specificity represent the absolute ranges at a prevalence of 2%.

and three fewer cervical cancer cases per million women screened. Correspondingly, the estimated effect of increasing specificity with 1% would be six fewer premature deliveries, 1.3 fewer major infections, and 3.3 fewer major bleeds per million women screened.

Using the estimated effects on downstream health outcomes and the values suggested previously, the guideline development group decided to accept a 4.5% decrease in specificity for every percentage increase in sensitivity (calculation in Appendix Table 1). This means that even if the lower limit of the CI of sensitivity (11% increase) were true, we would accept an increase in specificity of 50%. Because the entire CI of specificity is within this range, one would not rate down for imprecision in the specificity outcome. For the same reason, we would not rate down for imprecision in the sensitivity outcome. One should, however, also consider the uncertainty of the estimated downstream health outcomes on which we are basing the chosen range. Is it, for example, possible that the increased risk of premature delivery in treated women is 0.4% instead of the estimated 0.06%? If this is plausible, we would only accept an increase in specificity of 0.9% for every percentage increase in sensitivity (calculation in Appendix Table 2), and rating down for both imprecision and inconsistency for sensitivity and specificity would be warranted.

Just as for intervention effects, the fully contextualized ratings represent a sensitivity analysis addressing whether the test outcomes being considered (in this case sensitivity and specificity) are influential in altering the overall net benefit or harm.

8. Certainty ratings in the evidence to decision framework

As illustrated in Fig. 2 the ranges with different levels of contextualization will take into account one or several of the criteria in the evidence to decision framework [2]. For the noncontextualized ranges, only test accuracy is considered, whereas some level of consideration of the positive and negative health outcomes and values of these will be needed to set the partially contextualized ranges. For the fully contextualized ranges, all direct and downstream health outcomes are considered, as well as the balance of effects based on patient values. Depending on the perspective taken in the guideline, one could choose to also incorporate resource use, as well as issues of equity, acceptability, and feasibility when setting the fully contextualized ranges for sensitivity and specificity. For example, from a policy makers' perspective, one might want to include resources, such as further expensive testing in FPs or more expensive treatments due to delayed diagnosis in FNs.

9. Discussion

This paper illustrates the concepts of certainty of evidence applied to test accuracy. We show that defining ranges for the certainty ratings is important because the ranges chosen will affect the interpretation of the result and the degree of certainty presented. More empirical data are needed to inform approaches for defining ranges that would be most useful and to what degree different levels of modeling will affect the decisions being made.

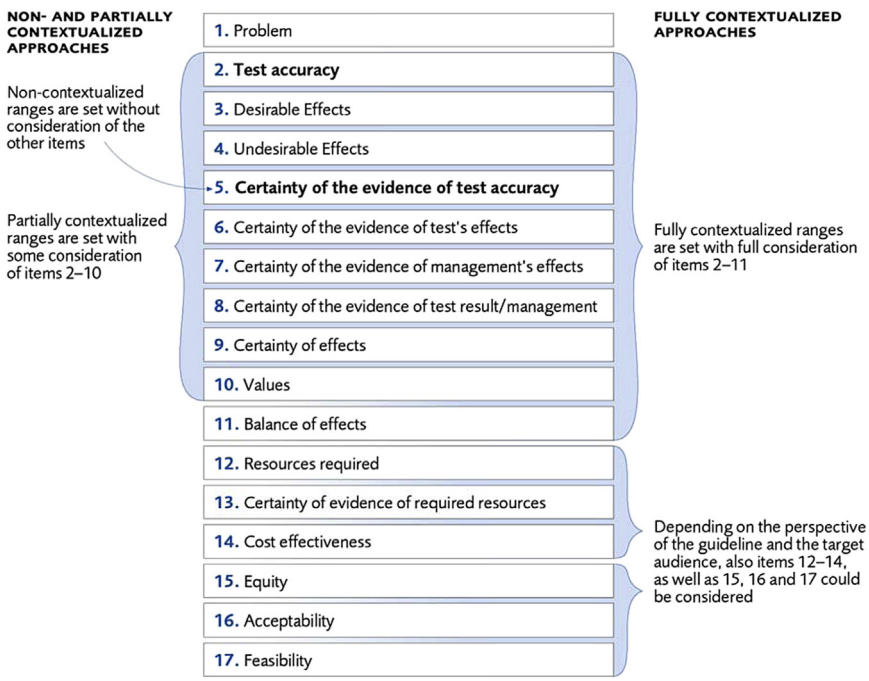

Fig. 2. The 17 items in the evidence to decision framework [2] and how the illustrated noncontextualized, partially contextualized, and fully contextualized approaches for setting ranges related to them.

Situations also exist that are complicated by issues related to research on tests. For example, primary research on test accuracy is historically seldom performed with direct comparisons between tests. Therefore, most often systematic review authors, health technology assessors, and guideline developers will not have access to primary studies directly comparing the accuracy of the relevant test strategies. Although this is starting to change, currently many decisions will have to be made based on indirect comparisons. Future studies are needed to inform how best to deal with these specific challenges.

Modeling of downstream health outcomes will inevitably include assumptions, which some review authors might feel reluctant to make. However, making decisions about tests will always require judgments regarding the importance of outcomes, although guideline panels or decision makers may not make their judgments explicit. An advantage of the fully contextualized approach for guideline development presented in this paper is the transparency of all assumptions made.

10. Conclusions

Previous work has shown that the certainty of evidence is a rating of our certainty that the true effect lies in a particular range. Although the examples related to intervention effects, previous guidance suggested that review authors specify the relevant thresholds underlying the certainty judgments. This guidance also applies to questions of diagnosis. In this conceptual paper, we have illustrated what the suggested approaches for defining ranges would mean when rating sensitivity and specificity in the context of systematic reviews, health technology assessments, or guidelines.

CRediT authorship contribution statement

Monica Hultcrantz: Conceptualization, Methodology, Investigation, Writing - original draft. **Reem A. Mustafa:** Conceptualization, Methodology, Investigation, Writing - original draft. **Mariska M.G. Leeflang:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft. **Valéry Lavergne:** Conceptualization, Methodology, Investigation, Writing - original draft. **Kelly Estrada-Orozco:** Conceptualization, Methodology, Investigation, Writing - original draft. **Mohammed T. Ansari:** Conceptualization, Methodology, Writing - original draft. **Ariel Izcovich:** Conceptualization, Writing - review & editing. **Jasvinder Singh:** Conceptualization, Writing - review & editing. **Lee Yee Chong:** Conceptualization, Writing - review & editing. **Anne Rutjes:** Conceptualization, Writing - review & editing. **Karen Steingart:** Conceptualization, Writing - review & editing. **Airton Stein:** Conceptualization, Writing - review & editing. **Nigar Sekercioglu:** Conceptualization, Writing - review & editing. **Ingrid Arevalo-Rodriguez:** Conceptualization, Writing - review & editing. **Rebecca L. Morgan:** Conceptualization, Writing - review & editing. **Gordon Guyatt:** Conceptualization, Writing - review & editing. **Patrick Bossuyt:** Conceptualization, Writing - review & editing. **Miranda W. Langendam:** Conceptualization, Methodology, Investigation, Writing - original draft. **Holger J. Schünemann:** Conceptualization, Methodology, Investigation, Writing - review & editing.

Acknowledgments

The authors would like to thank the colleagues who have contributed to the article during group discussions at GRADE Working Group meetings or discussions at the Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU). The authors also thank Dr Emelie Heintz for input on health economic modeling and Graphic Designer Elin Rye-Danjelsen for assisting with the figures, both at SBU.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.05.002.

References

[1] Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017;87:4−13.

[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. https://doi.org/10.1016/j.jclinepi.2019.02.003. pii: S0895-4356(17)31095-8.

[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 2008;336: 1106−10.

[5] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332:1089−92.

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

[7] Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013;158:544−54.

[8] 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.

[9] Mustafa RA, Wiercioch W, Ventresca M, Brozek J, Schunemann HJ, group DU-De. Decision making about healthcare-related tests and diagnostic test strategies. Paper 5: a qualitative study with experts suggests that test accuracy data alone is rarely sufficient for decision making. J Clin Epidemiol 2017;92:47−57.

[10] 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:e0134553.

[11] Mustafa RA, Wiercioch W, Arevalo-Rodriguez I, Cheung A, Prediger B, Ivanova L, et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 4: international guidelines show variability in their approaches. J Clin Epidemiol 2017;92:38−46.

[12] Pepe MS, Janes H, Li CI, Bossuyt PM, Feng Z, Hilden J. Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? Clin Chem 2016;62:737−42.

[13] Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29:E1−12.

[14] Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ 2011;343:d4684.

[15] 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.

[16] 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. Int J Gynaecol Obstet 2016;132(3):259−65.

[17] 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. Int J Gynaecol Obstet 2016;132(3):252−8.

[18] Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726−35.

[19] Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-risk of bias and indirectness. J Clin Epidemiol 2018;111:94−104.

[20] Zhang Y, Coello PA, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. J Clin Epidemiol 2018;111:83−93.

[21] Kuppermann M, Melnikow J, Slee C, Tancredi DJ, Kulasingam S, Birch S, et al. Preferences for surveillance strategies for women treated for high-grade precancerous cervical lesions. Gynecol Oncol 2010;118(2):108−15.

[22] Soergel P, Makowski L, Schippert C, Staboulidou I, Hille U, Hillemanns P. The cost efficiency of HPV vaccines is significantly underestimated due to omission of conisation-associated prematurity with neonatal mortality and morbidity. Hum Vaccin Immunother 2012;8(2):243−51.

[23] Wang K, Li H, Kwong WJ, Antman EM, Ruff CT, Giugliano RP, et al. Impact of spontaneous extracranial bleeding events on health state utility in patients with atrial fibrillation: results from the ENGAGE AF-TIMI 48 trial. J Am Heart Assoc 2017;6(8):1−7.