KCE PROCESS NOTE

GRADE FOR DIAGNOSIS

(*GCP*, HTA)

2017 www.kce.fgov.be

Version: 1.0, 2017-12-20

Contents

1. Introduction

1.1. What is GRADE FOR DIAGNOSIS?

Author(s): Lotty Hooft

Rob Scholten

Joan Vlayen

GRADE (Grading of Recommendations, Assessment, Development and Evaluation) offers a system for rating quality of evidence in systematic reviews and guidelines and grading strength of recommendations in guidelines. The system is originally designed for reviews and guidelines that examine alternative management strategies or interventions, which may include no intervention or current best management. It tries to offer a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines in health care and for carrying out the steps involved in developing recommendations (see 'What is GRADE' http://processbook.kce.fgov.be/node/107)

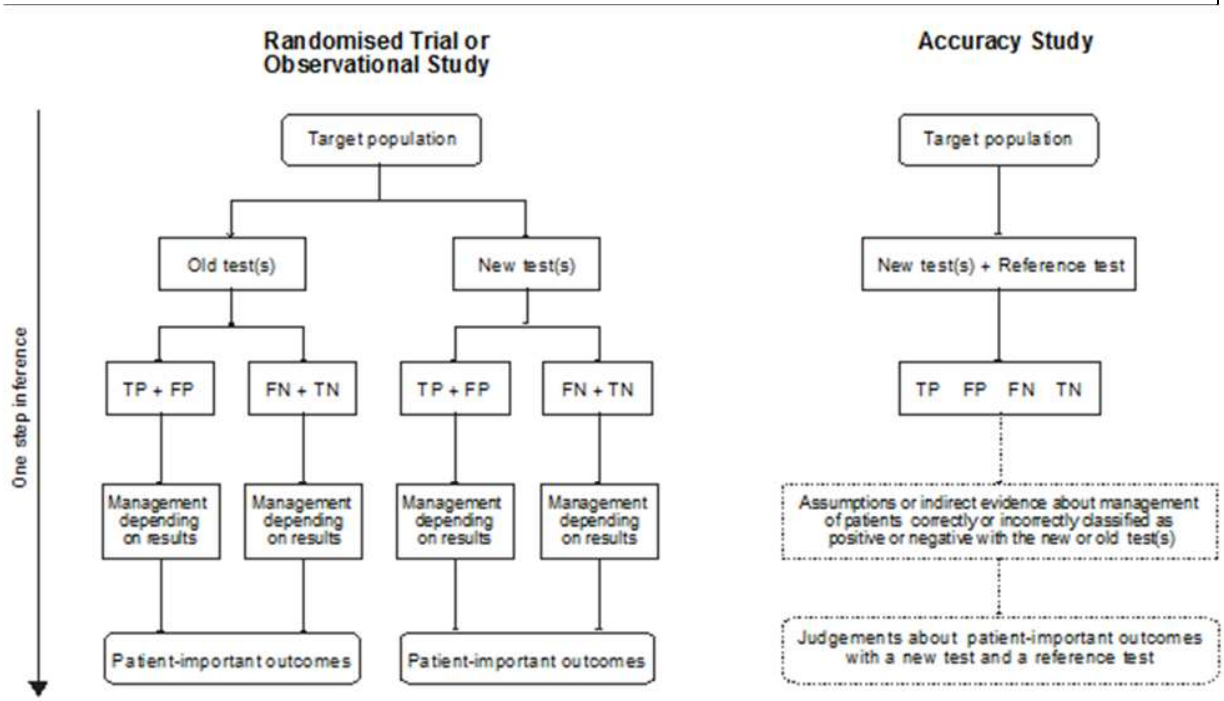
The GRADE approach can also be used for comprehensive and transparent rating of the quality of evidence in systematic reviews and for grading the strength of evidence-based recommendations about diagnostic tests or diagnostic strategies in clinical practice guidelines. Although this shares the basic principles of grading the quality of evidence and strength of recommendations for interventions, diagnostic questions present unique challenges. There are still a number of limitations and problems that are not entirely solved yet. However, there are some informative publications coming from the GRADE working group on this topic to guide authors of systematic reviews and guideline developers using GRADE to assess the quality of a body of evidence from diagnostic test accuracy (DTA) studies (Schünemann 2008, Brozek 2009, Hsu 2011, Schünemann 2016).

In the GRADE approach, also for diagnostic questions, the first essential step is specifying the patient-important outcomes (like mortality, morbidity, symptoms, and quality of life) before gathering and rating

the evidence, and grading the recommendations. However, diagnostic test research rarely focuses on patient important outcomes. Usually, when clinicians think about diagnostic tests, they focus on accuracy (sensitivity and specificity), i.e. how well the test classifies patients correctly as having or not having a disease. The underlying assumption is that obtaining a better idea of whether a target condition is present or absent will result in improved outcome. This is far from always the case and the degree to which this assumption holds varies a lot.

Recommendations about diagnostic tests should therefore consider whether the combination of establishing a diagnosis and treatment strategy as a whole will achieve a positive benefit/deficit ratio. This requires evidence comparing alternative test-treatment strategies focusing on patient important outcomes. A randomized clinical trial (RCT) is considered as the ideal research design for the evaluation of the effects of different test-treatment combinations on patient important outcomes (Figure 1).

Figure 1 (Source: GRADE Handbook - Chapter 7). Generic study designs that guideline developers can use to evaluate the impact of testing.



Two generic ways in which a test or diagnostic strategy can be evaluated. On the left, patients are randomized to a new test or to an old test and, depending on the results, receive the best available management (Bossuyt 2000). On the right, patients receive both: (one or more) new test(s) and reference test. Investigators can then calculate the accuracy of the new test(s) compared with the reference test (first step). To make judgments about the relation of new test accuracy to patient-important outcomes, one needs to make additional assumptions (relying on the information from subsequently or previously done studies) about successive management and likely outcomes of patients categorized with a new test or a reference test as either having or not having a target condition (second step)(Schünemann 2008).

TP: true positive, FP: false positive, TN, true negative, FN, false negative.

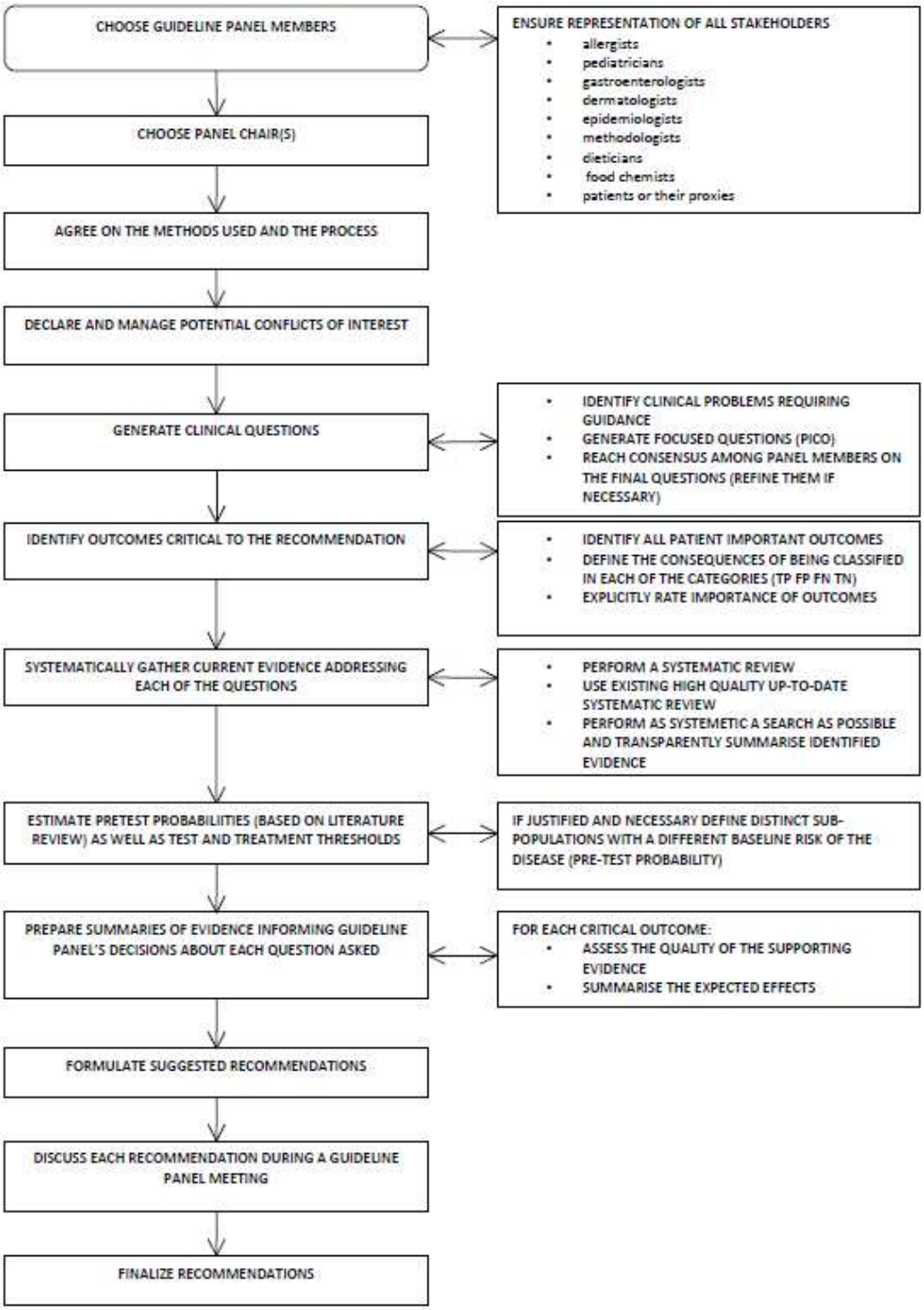
When diagnostic intervention studies – ideally RCTs but also nonrandomized studies – directly comparing the impact of alternative test-treatment strategies on patient-important outcomes are available, guideline

panels can use the GRADE approach as described for interventions studies (see http://processbook.kce.fgov.be/node/51). However, such studies require large sample sizes and rarely exist.

When studies providing direct evidence on the impact of tests on important outcomes are not available, guideline panels must focus on the performance of a test (diagnostic accuracy). Sometimes, considerations about accuracy and potential complications related to the test are sufficient for making a decision, especially in the case of comparative questions. In the case of comparative accuracy questions (when the new test is assessed against an existing diagnostic pathway), high quality evidence of test accuracy can be sufficient for making a decision or recommendation (see Chapter 3 'Deciding on important outcomes'). The GRADE working group has established criteria for assessing the quality of the evidence about test accuracy studies and making recommendations about tests in clinical guidelines (Schünemann 2008, Brozek 2009, Hsu 2011, Schünemann 2016). They also provided a general process to follow for developing clinical practice guideline on diagnostic tests (Figure 2).

If evidence on test accuracy measures is not sufficient for making a decision, guideline panels must consider to make inferences about the likely impact on patient-important outcomes. In addition to test accuracy measures, evidence or assumptions about the prevalence of the target condition, direct effects (benefits and harms) of the test, downstream effects of interventions, and prognosis of patients are needed. In particular, the downstream effects of the intervention(s) that follow based on the test results should be linked to the accuracy data. This requires judgments about the relationship between test results and patient-important consequences. The ideal approach is to have a fully developed model, with detailed assumptions and calculations for transparency. When detailed models or calculations are not available, guideline panels must make assumptions based on the perceived benefits and harms of the test, providing additional considerations that inform their judgments.

Figure 2. General process followed for developing clinical practice guideline on diagnostic tests (Hsu 2011).



For this reason, three types of evidence summaries (layers) can be used in the GRADE approach for diagnostic studies. First, there are evidence profiles that are based on diagnostic test accuracy alone (**layer 1**). They are the product of a DTA systematic review. The second type of evidence summary, includes information needed for healthcare decision-making, and includes DTA plus basic information about other features related to the test or test strategy facilitating decision making, such as direct benefits, harms or burden of the test and the number of inconclusive results (**layer 2**). The third format presents the effect of the test-treatment strategy on the patient-important outcomes, and includes explicit judgments about the consequences of tests (**layer 3**). This third layer requires that test results will be linked to management decisions. Evidence for each linkage might vary in quality. The linkages leading from changes in patient management decisions to patient important outcomes (**layer 3**) are of particular importance. Therefore, the value of a test or diagnostic strategy is often derived from their influence on treatment decisions and patient management.

Inferring from data on accuracy that a diagnostic test or strategy improves patient-important outcomes is preferably based on the availability of effective treatment. Even without an effective treatment, an accurate test may be beneficial if it reduces test related adverse effects or anxiety, or improves patients' and caregivers wellbeing through the prognostic information it imparts.

This example comes from the KCE Process Book: 7. GRADE and diagnostic testing; 7.1. Indirectness: test accuracy is a surrogate for outcomes important to patients.[1]

Example

The results of genetic testing for Huntington's chorea, an untreatable condition, may provide either welcome reassurance that a patient will not have the condition or the ability to plan for the future knowing that he or she will develop the condition. The ability to plan is analogous to an effective treatment, and the benefits of planning need to be balanced against the downsides of receiving an early diagnosis.

[1] http://processbook.kce.fgov.be/node/151

1.2. Limitations of GRADE

See KCE process book GRADE System (GCP, HTA)[2]

No specific limitations of GRADE for Diagnosis are provided.

1.3. Steps in the process

See KCE process book GRADE System (GCP, HTA

Author(s):

Jo.Robays

Joan.Vlayen

GRADE for diagnosis includes the following steps:

- Ask a specific healthcare question to be answered by a recommendation;
    - establishing the purpose of a test
    - establishing the role of a test: triage, replace a current test or add-on
    - establishing the reference test
- Identify all important outcomes for this healthcare question;
- Judge the relative importance of outcomes;
- Summarize all relevant (direct and indirect) evidence;
    - For Layer 1: evidence on the performance of a test,
    - For Layer 2: evidence on the performance of a test and assumptions on direct benefits, harms or burden of the test,

[2] http://processbook.kce.fgov.be/node/106

        - Inconclusive test results
        - Complications of a test
    - For Layer 3: evidence needed for Layer 2 and evidence on downstream consequences of the test (linked evidence on natural course of the condition and desirable and undesirable effects of clinical management).
- Grade the quality of all types of underlying evidence;
- Decide on the overall confidence in estimates of effects;
- Include judgments about the underlying values and preferences related to the management options and outcomes (Evidence to Decision making);
- Decide on the balance of desirable and undesirable effects;
- Decide on the balance of net benefits and cost;
- Grade the strength of recommendation;
- Formulate a recommendation;
- Implement and evaluate.

2. Framing the question

Author(s): Lotty Hooft

Rob Scholten

Applying GRADE for diagnosis begins with formulating appropriate clinical questions using 'PICO'[3] or another structured format (Hsu 2011). Formulating clinical questions requires clearly establishing the role or purpose of a test or diagnostic strategy. The format of the question asked by authors of systematic reviews or guideline developers follows the same principles as the format for management questions: pertaining to a defined population (P) for whom the test or diagnostic strategy (I) is being considered in relation to a comparison test or diagnostic strategy (C) according to defined patient outcomes of interest (O).

First the systematic review authors or guideline panels determine the **population** of interest (i.e., those in whom the diagnosis is uncertain). An **index** test (i.e., the test of interest or a 'new' test) can play one of three roles in the existing diagnostic pathway: act as a triage (to minimize use of invasive or expensive tests), replace a current test (to eliminate tests with worse test performance compared to a current test, greater burden, invasiveness, or cost), or add-on (to enhance accuracy of a diagnosis beyond current test) (Bossuyt 2006). A challenging situation can occur when panel members indicate that they are interested in the index tests as a replacement for the reference standard (the best available test able to distinguish diseased from non-diseased). In situations where the reference standard is imperfect, the estimates of sensitivity and specificity are biased. There are several methodological options for how to summarize diagnostic accuracy in the absence of a gold standard, however, these discussions go beyond the scope of this GRADE for diagnosis process book.

[3] See also http://processbook.kce.fgov.be/node/110

Guidelines often need an additional specification of the setting in which the guideline will be implemented. For instance, guidelines intended for resource-rich environments will often be inapplicable in resource-poor environments. Furthermore, in some cases it may be necessary to specify if the guideline needs to be implemented in an inpatient or an outpatient setting.

To ensure that guideline panels can develop informed recommendations about diagnostic tests, more emphasis should be placed on group processes, including question formulation, defining patient-important outcomes for diagnostic tests, and summarizing evidence. Explicit consideration of concepts of diagnosis from evidence-based medicine, such as pre-test probability and treatment threshold, is required to facilitate the work of a guideline panel and to formulate implementable recommendations. Recommendations may differ across subgroups of patients, as also evidence quality may differ across subgroups. Thus, guideline panels often should define separate questions (and produce separate evidence summaries) for different settings or subgroups of patients.

Possible formats for diagnostic research questions may be the following:

- Should TEST A vs. TEST B be used in SOME PATIENTS/POPULATION?

- Should TEST A vs. TEST B be used for SOME PURPOSE?

In practice, however, these formats will usually be combined. In addition, the use of one test may be compared with no testing, which means that only one test might be mentioned in the research question.

Examples

- Should frozen section analysis be used to diagnose malignant ovarian tumours in women suspected of early-stage ovarian cancer?

- Should Exercise ECG vs. CT coronary angiography be used for diagnosing coronary stenosis 50% or more in patients with stable angina pectoris suspected of coronary disease?

3. Deciding on important outcomes

3.1. General approach

*See **KCE process book GRADE System (GCP, HTA) and GRADE Handbook**[4]*

Author(s): Lotty Hooft

Rob Scholten

Joan Vlayen

GRADE specifies three categories of outcomes according to their importance. Guideline developers must, and authors of systematic reviews are strongly encouraged to specify all potential patient-important outcomes as the first step in their endeavour. If a test or diagnostic strategy fails to improve patient-important outcomes there is no reason to use it, whatever its accuracy. The guideline panel should classify outcomes as:

- Critical;
- Important, but not critical;
- Of limited importance.

The first two classes of outcomes will bear on guideline recommendations; the third may or may not. Ranking outcomes by their relative importance can help to focus attention on those outcomes that are considered most important, and help to resolve or clarify disagreements. GRADE recommends to focus on a maximum of 7 critical and/or important outcomes (similar to GRADE for management decisions).

[4] http://processbook.kce.fgov.be/node/112

GRADE recommends a systematic classification of the importance of outcomes to patients on a 9-point scale as: not important (score 1–3), important, but not critical 4–6, or critical 7–9 to a decision.

In the GRADE system for diagnosis, valid accuracy studies can provide high quality evidence of test accuracy (layer 1). In the case of comparative accuracy questions (when the new test is assessed against an existing diagnostic pathway), high quality evidence of diagnostic accuracy studies can be sufficient for making a decision or recommendation. In these situations, only layer 1 of the GRADE system for diagnosis is needed. For example, when the new test is as sensitive and as specific as the old test and the new test has advantages over the old test. This is a straightforward example, however, in many situations this might be less clear. Knowledge is needed about the comparability of the downstream consequences of the new and the old test and the type of cases (same or different) detected by both test. . Key questions to consider helpful in determining whether to focus exclusively on accuracy studies (layer 1) are (Samson 2012):

1. Are the extra cases detected by the new, more sensitive test similarly responsive to treatment as are those identified by the older test?;
2. Are trials available that selected patients using the new test?;
3. Do trials assess whether the new test results predict response?;
4. If available trials selected only patients assessed with the old test, do extra cases identified with the new test represent the same spectrum or disease subtypes as trial participants?;
5. Are tests' cases subsequently confirmed by same reference standard?;
6. Does the new test change the definition or spectrum of disease (e.g., by finding disease at an earlier stage)?;
7. Is there heterogeneity of test accuracy and treatment effect (i.e., do accuracy and treatment effects vary sufficiently according to levels of a patient characteristic to change the comparison of the old and new test)?

However, evidence on test accuracy is often not adequate and is only considered as a surrogate for patient-important outcomes. Similar to treatment management strategies, patient-important outcomes

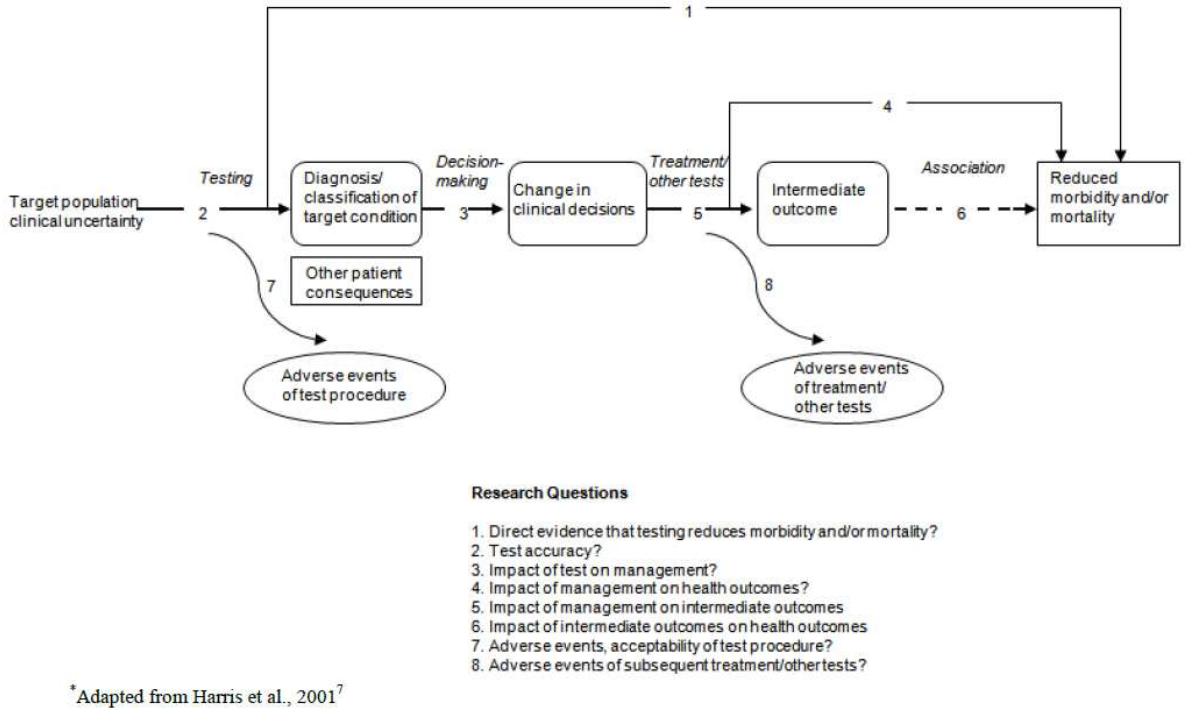
may include survival (mortality), clinical events (e.g. stroke or myocardial infarction), patient-reported outcomes (e.g. specific symptoms, quality of life), adverse events, burden (e.g. demands on caregivers, frequency of tests, restrictions on lifestyle) and economic outcomes (e.g. cost and resource use). It is critical to identify both outcomes related to effectiveness as well as outcomes related to adverse effects/harm.

GRADE provides a structured framework for diagnosis (Hsu 2011) that considers the following outcomes:

- The patient important consequences of correctly classified patients (true positive (TP) and true negative (TN))
- The patient important consequences of being incorrectly classified patients (false positive (FP), or false negative (FN));
- The consequences of inconclusive results;
- Complications of a new test and a reference standard; and resource use (cost).

Hsu 2011 wrote: ‘Correct classification is usually associated with benefits or a reduction in adverse outcomes, while incorrect classification is associated with worse consequences (harms), including failure to treat and potentially reduce burden of disease. A guideline panel needs to evaluate whether the benefits of a correct classification (TP and TN) outweigh the potential harms of an incorrect classification (FP and FN). However, the benefits and harms follow from subsequent action and are determined by probabilities of outcome occurrence and the importance of these outcomes to patients (e.g. mortality, morbidity, symptoms).

Figure 3. Analytic framework to test evaluation (Samson 2012).



*Adapted from Harris et al., 2001[7]

Review authors should consider which outcomes they want to assess (Figure 3) and how these outcomes should be measured, both in terms of the type of scale likely to be used and the timing of measurement. Linking of the different types of evidence needed, requires judgments about the relationship between test results, treatment decisions and patient-important consequences.

3.2. Perspective of outcomes

See KCE process book GRADE System (GCP, HTA)[5]

Author(s):

Jo.Robays

[5] http://processbook.kce.fgov.be/node/113

Joan.Vlayen

Different audiences are likely to have **different** perspectives on the importance of outcomes.

The importance of outcomes is likely to vary within and across cultures or when considered from the perspective of patients, clinicians or policy-makers. It is essential to take cultural diversity into account when deciding on relative importance of outcomes, particularly when developing recommendations for an international audience. Guideline panels should also decide what perspective they are taking. Guideline panels may also choose to take the perspective of the society as a whole (e.g. a guideline panel developing recommendations about pharmacological management of bacterial sinusitis may take the patient perspective when considering health outcomes, but also a society perspective when considering antimicrobial resistance to specific drugs).

3.3. Before and after literature review

See KCE process book GRADE System (GCP, HTA)[6]

No specific guidance for GRADE for Diagnosis is provided.

3.4. Implications of the classification

See KCE process book GRADE System (GCP, HTA)[7]

No specific guidance for GRADE for Diagnosis is provided.

3.5. Expert involvement

See KCE process book GRADE System (GCP, HTA)[8]

[6] http://processbook.kce.fgov.be/node/114

[7] http://processbook.kce.fgov.be/node/115

[8] http://processbook.kce.fgov.be/node/116

No specific guidance for GRADE for Diagnosis is provided.

3.6. Clinical decision threshold and minimally important difference

Author(s): Lotty Hooft

Rob Scholten

To enable judgment of the effects of testing (and thus the importance of the various outcome categories of a test), summary estimates of sensitivity and specificity are not very informative. In GRADE for management decisions, the results of a systematic review (usually expressed as summary risk ratios or summary odds ratios) are translated into a format with natural numbers that can be more readily understood by end-users of the review. The same procedure is done by review authors with GRADE for Diagnosis. The summary estimates of sensitivity and specificity (and their 95% confidence intervals) are applied to a hypothetical population of – usually – 1000 patients that are suspected of having the target condition at hand and will undergo the test of which the accuracy was assessed.

The starting point for construction of such a summary 2*2 Table is the selection of an estimate of the prior probability (prevalence) of the target condition in the population to which the test will be applied. If this prevalence is known (e.g. from the literature or registries), this prevalence would be a good choice. One should, however, refrain from selecting a prevalence value outside the range of the included studies. Such an extrapolation is not allowed, because no data exist to assess whether the results of the meta-analysis would be applicable to that extrapolated value.

In most instances, however, such information from the literature will not be available. In that case one could take the median of the prevalence of the target condition of the studies included in the review as a

starting point for constructing the required 2*2 Table. To allow for variation (or imprecision) of the prevalence estimate, the 25th and 75th percentiles (1st and 3rd quartiles) could also be included.

An example of how to construct such a summary 2*2 Table is presented in Box 1. Figure 4 presents the same example in the format of a GRADE layer 1 SoF Table, as constructed with the online guideline development tool.

Examples.

BOX 1. *Review question: Should frozen section analysis be used for diagnosing malignant ovarian tumours in women suspected of early-stage ovarian cancer?*

Summary sensitivity: 90.3% (95% CI 88.0% to 92.2%)

Summary specificity: 99.5% (95% CI 99.1% to 99.7%)

Assumed prevalence of the target condition: 29% (= median prevalence of the 37 included studies)

Number of patients tested: 1000

Number of patients with the target condition: 0.29*1000 = 290

Number of TPs: 0.903*290 = 262

Number of TNs: 0.995*710 = 706

Resulting 2*2 Table:

	Target condition		
	Present	Absent	Total
Index test +	262	4	266

Index test -	28	706	790
Total	290	710	1000

Figure 4. *Should frozen section analysis be used for diagnosing malignant ovarian tumours in women suspected of early-stage ovarian cancer?*

Patient or population : women suspected of early-stage ovarian cancer

New test: frozen section analysis

Reference test: laparotomy | **Threshold :** N/A

Pooled sensitivity : 0.903 (95% CI: 0.880 to 0.922) | **Pooled specificity** : 0.995 (95% CI: 0.991 to 0.997)

Test result	Number of results per 1.000 patients tested (95% CI) Prevalence 29% (median)	Number of participants (studies)	Quality of the Evidence (GRADE)
True positives	**262** (255 to 267)	3096 (37)	⨁⨁⨁◯ **MODERATE** [a]
False negatives	**28** (23 to 35)		
True negatives	**706** (704 to 708)	7431 (37)	⨁⨁⨁◯ **MODERATE** [a]
False positives	**4** (2 to 6)		

CI: Confidence interval

Explanations

a. Unclear risk of bias in the majority of studies; most studies retrospective.

4. Summarizing the evidence

See GRADE Handbook[9]

Author(s): Lotty Hooft

Rob Scholten

Miranda Langendam

Mariska Tuut

A guideline panel should base its recommendation(s) on the best available body of evidence related to the health care question and preferably to all patient important outcomes. GRADE recommends that systematic reviews should form the basis for making health care recommendations. The evidence collected from systematic reviews or health technology assessments is used to produce GRADE evidence profiles and summary of findings tables, for which GRADEpro GDT can be used (www.gradepro.org). This is an easy to use all-in-one web solution for summarizing and presenting information for health care decision making and the development of clinical practice guidelines.

Evaluating diagnostic tests or strategies in terms of impact on patient important outcomes involves linking different components of the evidence. The components are:

a. Test accuracy (layer 1)
b. Direct benefits, harms or burden of the test (layer 2)
c. Downstream consequences of the test (layer 3)
d. Impact of the test on management decisions (layer 3)

a) Test accuracy

[9] http://gdt.guidelinedevelopment.org/app/handbook/handbook.html

The test accuracy determines the number of TP, TN, FP and FN per 1,000 patients tested, for a specific pre-test probability. The estimated accuracy of tests or diagnostic strategies, preferably based on a meta-analysis, comes from systematic reviews of diagnostic test accuracy studies.

In GRADEpro GDT this is called:: *How accurate is the test?*

b) Direct benefits, adverse effects or burden of the test

Conducting the test can have direct negative effects, for example if a test is invasive (e.g. risk of major bleeding) or can cause allergic reactions (risk of septic shock). An example of a positive direct effect is removal of the blockage by the radiographic contrast (dye) used in a hysterosalpingogram (HSG), a test of female fertility potential. Often this evidence comes from test accuracy studies, but it may come from other studies that evaluated the use of the test.

c) Downstream consequences of the test (linked evidence)

I. Natural course of the condition

The natural course of the condition informs about the consequences of a FN test result. For example, the health consequences of having a FN result are more serious if the condition is progressive. The evidence usually comes from prognostic studies that estimate the risk of developing the outcomes without treatment.

II. Desirable and undesirable effects (effectiveness) of clinical management

The effectiveness of the treatment informs about the consequences of a positive test result. Effective treatment will lead to health benefit for the TP. Those with a FP result however, will not benefit from the treatment, but may experience the adverse effects or complications. The evidence comes from intervention studies. In clinical guidelines, the treatment effectiveness question may be part of the guideline as well, and can be linked to the diagnostic question.

An example of assessing the patient-important consequences of being classified into TP, TN, FP and FN categories can be found in Table XX (based on Kapur 2017).

Example of linking patient important outcomes to true positive (TP), false positive (FP), false negative (FN) and true negative (TN) test results (adapted from Kapur 2017)	
Question:	In adult patients with suspected obstructive sleep apnea syndrome (OSA), does home sleep apnea testing (HSAT) accurately diagnose OSA, improve clinical outcomes and improve efficiency of care compared to polysomnography (PSG)?
Population:	Adult patients with suspected OSA
Index test:	HSAT
Comparator:	PSG
Outcomes	**Consequences**
TP:	Effective treatment and improved QOL Side-effects of therapy Improvement in comorbid conditions (e.g., hypertension) Reduced risk of CV events Reduced risk of post-CV events Reduced risk of motor vehicle accident (MVA) Reduced overall health costs
TN:	Confirmation of absence of OSA Possible repeat testing if patient deemed at high risk for OSA Psychological relief Consideration of alternative causes for symptoms Saves time and resources Focused treatment on true cause of symptoms
FP:	Unnecessary treatment and utilization of resources Increased costs due to treatment Psychological distress

	Delay in diagnosis of true condition Side-effects of therapy
FN:	Absence of necessary treatment Reduced QOL Psychological distress Possible repeat testing if patient deemed at high risk for OSA Risk of motor vehicle accident (MVA) Risk of hypertension Risk of CV events Post-MI events Post-stroke events Death Increased costs and utilization of resources due to other condition(s)
Inconclusive results:	Proportion of patients in whom HSAT doesn't provide a result
Complications of the test:	None known
Costs:*	Not addressed

* Not in the current Layer 2.

In GRADEpro GDT the patient-important outcomes and the treatment that follows from a positive test result can be listed in the Question part of the Recommendations section (per diagnostic question).

d) Link between the test results and the management decisions: will a given test result be managed according to that result?

If a given test result will not lead to (adequate) clinical management, for example for practical or organizational reasons or patient or clinician preferences, the impact of the test on patient important

outcomes will not be optimal. The evidence may come from different types of study design, including qualitative research.

In GRADEpro GDT this is called:

How substantial are the desirable anticipated effects?

This refers to the downstream consequences of having a TN or TP test result and/or direct benefit of test.

How substantial are the undesirable anticipated effects? (FN, FP and direct harm or burden of test)

This refers to the test accuracy, more specifically the natural frequencies of FN and FP (e.g. per 1,000 patients tested), the downstream consequences of having a FN or FP test result and/or direct harm or burden of test.

Ideally, guideline developers will evaluate and rate a body of evidence for each of these four components (accuracy, direct effects of the test, downstream consequences and link between test results and management decisions) by means of systematic reviews. If this is not possible, for example because of limited resources, this should be stated.

Linking the evidence: from test results to patient important outcomes

Linking the test results (TP, FP, TN, FN) to the downstream consequences (patient important outcomes) can be done informally ("back of the envelop approach") or formally in a decision model. In the informal approach the guideline panel discusses and describes the health consequences of having a TP, TN, FP (unnecessary treated) or FN (missed or delayed diagnosis) result (for an example see Hsu 2011).

Formal modeling is a quantitative approach. An example of this approach can be found in the WHO Guidelines for screening and treatment of precancerous lesions for cervical cancer prevention (World Health Organisation 2013, Santesso 2016). In this guideline, the effect of different screen-treatment strategies on patient important outcomes was estimated. The modeling resulted in estimates for the

number of cervical cancer deaths per 1,000,000 women screened ranging between 20 and 88 for the different screen-treatment options and 250 in case of no screening.

5. Rating the quality of evidence

5.1. Introduction

See GRADE Handhook[10]

Author(s): Lotty Hooft

Rob Scholten

As in GRADE for management decisions studies, GRADE for Diagnosis specifies four quality categories (high, moderate, low, and very low) that apply to a body of evidence, but not to individual studies. GRADE for Diagnosis also uses the same factors (Limitations in study methods ('Risk of bias'), Indirectness, Inconsistency, Imprecision, and probability of Publication bias) to determine the quality of evidence.

Quality level	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect
Moderate	We are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect
Very low	We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of the effect

[10] http://gdt.guidelinedevelopment.org/app/handbook/handbook.html

When randomized trials as described in Figure 1 are available, the approach to assessing the confidence in effect estimates (quality of evidence) for interventions should be used. When such direct data on patient-important outcomes are lacking, and the body of evidence is derived from DTA studies, the next step is to rate the quality of the evidence for the different layers of evidence (layer 1, 2, and 3) to judge the overall certainty of the underlying evidence about the effect of the test:

a) Quality of the evidence of test accuracy (layer 1);

b) Quality of the evidence of direct benefits, adverse effects or burden of the test (layer 2);

c) Quality of the evidence of the natural course of the condition and the effects of clinical management guided by the test results (layer 3);

d) Certainty of the link between test results and management decisions (layer 3);

Guideline panels have to determine the overall quality of evidence across all the outcomes that are essential to the recommendation they make. The strength of a recommendation usually depends on evidence regarding not just one, but a number of patient-important outcomes, and on the quality of (the different layers of) evidence for each of these outcomes.

When determining the overall quality of evidence across outcomes (also see Chapter 3):

- Consider only those outcomes that are deemed critical;
- If the quality of evidence differs across critical outcomes and outcomes point in different directions — towards benefit and towards harm — the lowest quality of evidence for any of the critical outcomes determines the overall quality of evidence;
- If all outcomes point in the same direction — towards either benefit or harm — the highest quality of evidence for a critical outcome, that by itself would suffice to recommend an intervention, determines the overall quality of evidence. However, if the balance of the benefits and harms is uncertain, the grade of the critical outcome with the lowest quality grading should be assigned.

However, it is important to emphasize that GRADE warns against applying downgrading in a too mechanistic way and to leave room for judgment. Although GRADE suggests the initial separate consideration of five categories for rating down the quality of evidence, with a yes/no decision in each case, the final rating of overall evidence quality occurs in a continuum of confidence in the validity, precision, consistency, and applicability of the estimates. Fundamentally, the assessment of evidence quality remains a subjective process, and GRADE should not be seen as obviating the need for or minimizing the importance of judgment. As repeatedly stressed, the use of GRADE will not guarantee consistency in assessment, whether it is of the quality of evidence or of the strength of recommendation. There will be cases in which competent reviewers will have honest and legitimate disagreement about the interpretation of evidence. In such cases, the merit of GRADE is that it provides a framework that guides one through the critical components of this assessment and an approach to analysis and communication that encourages transparency and an explicit accounting of the judgments involved.

5.2. Quality of evidence for test accuracy

Author(s): Lotty Hooft

Rob Scholten

In the context of a DTA systematic review, quality reflects our confidence that the accuracy estimates are correct. Factors that decrease the quality of evidence for studies of diagnostic accuracy (and how they differ from evidence for intervention studies), are presented in Table 1 coming from the GRADE Handbook.

Table 1. Factors that decrease the quality of evidence for studies of diagnostic accuracy (GRADE Handbook)

Factors that determine and can decrease the quality of evidence	**Explanations and how the factor may differ from the quality of evidence for other interventions**
Study design	Different criteria for accuracy studies Cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard (best possible alternative test strategy) are considered high quality and can move to moderate, low or very low depending on other factors.
Risk of bias (limitations in study design and execution)	Different criteria for accuracy studies 1. Representativeness of the population that was intended to be sampled. 2. Independent comparison with the best alternative test strategy. 3. All enrolled patients should receive the new test and the best alternative test strategy. 4. Diagnostic uncertainty should be given. 5. Is the reference standard likely to correctly classify the target condition?
Indirectness Patient population, diagnostic test, comparison test and indirect comparisons of tests	Similar criteria The quality of evidence can be lowered if there are important differences between the populations studied and those for whom the recommendation is intended (in prior testing, the spectrum of disease or co-morbidity); if there are important differences in the tests studied and the diagnostic expertise of those applying them in the studies compared to the settings for which the recommendations are intended; or if the tests being compared are each compared to a reference (gold) standard in different studies and not directly compared in the same studies. Similar criteria

	Panels assessing diagnostic tests often face an absence of direct evidence about impact on patient-important outcomes. They must make deductions from diagnostic test studies about the balance between the presumed influences on patient-important outcomes of any differences in true and false positives and true and false negatives in relationship to test complications and costs. Therefore, accuracy studies typically provide low quality evidence for making recommendations due to indirectness of the outcomes, similar to surrogate outcomes for treatments.
Important Inconsistency in study results	Similar criteria For accuracy studies unexplained inconsistency in sensitivity, specificity or likelihood ratios (rather than relative risks or mean differences) can lower the quality of evidence.
Imprecise evidence	Similar criteria For accuracy studies wide confidence intervals for estimates of test accuracy, or true and false positive and negative rates can lower the quality of evidence.
High probability of Publication bias	Similar criteria A high risk of publication bias (e.g., evidence only from small studies supporting a new test, or asymmetry in a funnel plot) can lower the quality of evidence.
Upgrading for dose effect, large effects residual plausible bias and confounding	Similar criteria For all of these factors, methods have not been properly developed. However, determining a dose effect (e.g., increasing levels of anticoagulation measured by INR increase the likelihood for vitamin K deficiency or vitamin K antagonists). A very large likelihood of disease (not of patient-important outcomes) associated with test results may increase the quality evidence. However, there is some disagreement if and how dose effects play a role in assessing the quality of evidence in DTA studies.

5.2.1. Study limitations (Risk of bias)

Author(s): Lotty Hooft

Rob Scholten

For the assessment of the methodological quality of individual DTA studies, the most frequently used tool is QUADAS-2 (Whiting 2011)[11]. QUADAS-2 addresses not only the risk of bias (methodological quality) of individual studies, but also the applicability of the studies to the clinical situation at hand. The results of the risk of bias part of QUADAS can be used for assessing study limitations across a body of evidence (like in GRADE for management decisions), whereas the applicability part can be used for assessing the directness of a body of evidence to the PICO question at hand (see next paragraph).

For the assessment of risk of bias QUADAS discriminates between four domains: patient selection, index test, reference standard and flow and timing.

Study limitations in DTA studies, as identified by GRADE, should be assessed separately for patients with the target condition (sensitivity estimates; TP and FN) and those without the target condition (specificity estimates TN and FP). These could be the following (rate down with one or two levels):

- Patient selection: lack of including a consecutive sample (or a random sample) of patients or the use of a case-control design for assessing the DTA of a test;
- Index test: interpreting the index test with knowledge of the results of the reference standard or not having used a pre-specified threshold for positivity of the index test;
- Reference standard: having used a reference standard that may not classify patients correctly as having or not having the target condition or interpretation of the reference standard with knowledge of the results of the index test;

[11] See also http://processbook.kce.fgov.be/node/155

- Flow and timing: inappropriate interval between conducting the index test and the reference standard, applying the reference standard to a subset of patients only (partial verification), not having applied the same reference standard to all patients (differential verification) or not having included all patients in the analysis.

Moving from risk of bias of individual studies to a judgment about rating down for study limitations across a body of DTA evidence mainly relies on judgment and is challenging. Many studies of test accuracy suffered very serious methodological limitations that warranted downgrading the quality of evidence by two levels; from high, through moderate, to low (Hsu 2011).So in principle, this judgment is similar to that of GRADE for management decisions (reference to paragraph 5.2.2. of the current KCE process book[12]), but it must be emphasized that the relationship between the various risk of bias items and the occurrence of true bias in DTA studies is still not that clear for the DTA domain.

In addition, due to incomplete reporting of DTA studies, which, unfortunately, is still very prevalent, many items might have been scored 'Unclear', hampering a sound judgment of the risk of bias. So, a judgment must be made about whether or not bias is expected for the estimated summary estimates of both sensitivity and specificity due to study limitations across studies.

5.2.2. Indirectness

Author(s): Lotty Hooft

Rob Scholten

In the DTA domain, indirectness can be defined on two different levels.

The first level of indirectness applies to the **directness (or applicability) of the patients**, **index tests and reference standards** that were assessed in the included studies, to those of the clinical question

[12] http://processbook.kce.fgov.be/node/128

(PICO) at hand. For this level the results of the three applicability domains of QUADAS-2 can be used: patient selection, index test and reference standard.

As before, these categories must be assessed separately for patients with the target condition (sensitivity estimates) and those without the target condition (specificity estimates).

The second level applies to the **comparison of index tests**. When the clinical question addresses the choice between test A and test B, most valid and unbiased evidence will come from studies that directly compare those tests (i.e. head-to-head comparisons). Such comparisons may come from studies that assess the DTA of both tests in all patients (paired design), or from randomised comparisons of tests, in which patients are randomised to either test A or to test B. In both cases all patients have been verified by the reference standard.

Usually, however, the body of evidence will consist of a set of studies that addressed test A and another set of studies that addressed test B. A comparison of the DTA of test A and B will then be indirect and, therefore, less trustworthy (like indirect comparisons in intervention research). In such cases one must consider downgrading for indirectness due to indirect comparisons.

However, studies that directly compare two tests could have led to a highly selected patient population that doesn't necessarily reflect the patients that one will see in daily clinical practice for whom the tests are meant. An example of this is a systematic review that compared the accuracy of exercise ECG with CT coronary angiography to diagnose coronary stenosis 50% or more in patients with stable angina pectoris suspected of coronary disease (Nielsen 2015). Amongst the inclusion criteria was the requirement that studies had to compare the two index tests directly. This criterion, however, has major implications for the type of patients that were selected in those studies. The ECG exercise test will usually be done in patients with less severe stages of coronary disease, whereas CT coronary angiography will be used in more advanced cases. This has led to major applicability concerns and to less trustful accuracy estimates.

5.2.3. Inconsistency

Author(s): Lotty Hooft

Rob Scholten

Analogous to the therapeutic domain the quality of the evidence should be downgraded by one or two levels (depending on the magnitude of the inconsistency in the results) when heterogeneity exists for which investigators were not able to identify a plausible explanation (reference to paragraph 5.3. of the current KCE process book). In meta-analysis of DTA studies, however, heterogeneity (or in GRADE terminology 'inconsistency') is the rule rather than the exception. This inconsistency can be caused by the inclusion of different study populations or differences in the use and/or definition of the index tests or reference standards (clinical heterogeneity) or poor methodology (methodological heterogeneity). Unlike in meta-analyses of intervention studies, the usual statistical assessments (Cochran's Chi-square test for homogeneity and I-square for the assessment of the percentage of the total variation caused by differences between studies (between-study variation)) are not very useful in the DTA domain. Assessment of inconsistency will mainly have to rely on visual inspection of the paired forest plots of sensitivity and specificity. Questions to consider are "Are the individual point estimates more or less the same?" and, more importantly, "Do confidence intervals sufficiently overlap?". A graph of the sensitivity and specificity pairs in receiver operating characteristic (ROC)-space is also very helpful to assess homogeneity ("Do the sensitivity-specificity pairs cluster sufficiently or are the spread all over the ROC space?"). In addition, the size of the 95% confidence ellipse around the pooled point estimates of sensitivity and specificity can help in the assessment of inconsistency. Finally, the size of the 95% prediction ellipse (if presented) can assist in this assessment. This ellipse indicates the region of which it's 95% likely that the estimate of sensitivity and specificity of a future study will lie. The larger the amount of heterogeneity across studies, the larger these ellipses will be.

Because in DTA meta-analyses mainly random effects models are used, the size of the two above-mentioned ellipses is not only influenced by inconsistency (heterogeneity between studies), but also by imprecision due to small study sizes (like in fixed effect models). Thus, a large 95% confidence ellipse

can be caused by inconsistency, imprecision or both (and one should not downgrade twice: see next paragraph).

5.2.4. Imprecision

Author(s): Lotty Hooft

Rob Scholten

How to apply GRADE for the assessment of imprecision of a body of DTA evidence, is still work in progress. Like in GRADE for management decisions, one could look at the width of the 95% confidence intervals of the pooled estimates of sensitivity and specificity, and assess whether they cross a certain clinically acceptable lower limit (for which we would downgrade), or not. However, it is not straightforward to define those lower limits, because for clinical decisions the consequences of testing are judged on the positive and negative predictive values of tests, which depend on the prevalence of the target condition. Therefore, a better option is to base this judgment on the 95% confidence intervals around the two categories of test results that have the most important consequences for patients: the number of FPs and the number of FNs. These 95% confidence intervals are derived from the 95% confidence intervals of the summary sensitivity and specificity and are presented in a SoF Table, which includes the summary 2*2 Table in a hypothetical population of 1000 tested.

5.2.5. Publication bias

Author(s): Lotty Hooft

Rob Scholten

High risk of publication bias can lower quality of evidence, mainly because studies with statistically significant results are published more rapidly than those without. However, if reporting bias exists in the field of diagnostic accuracy studies and whether the underlying mechanisms are similar, is still work in progress. Time to publication of diagnostic accuracy studies with promising results about the performance

of tests seems to are published more rapidly compared to those reporting lower estimates (Korevaar 2016).

Statistical methods for investigating publication bias (small-study effect) in test accuracy studies differ from those used in intervention studies, because of the different nature of diagnostic and intervention questions (Deeks 2005). On the other hand, small-study effects and time trends do not seem to be as pronounced in meta-analyses of test accuracy studies as they are in meta-analyses of randomized trials, although larger studies tend to report higher sensitivity in the field of diagnostic imaging studies (Korevaar 2016). Nevertheless, the majority of DTA review authors investigate publication bias. They mainly use suboptimal methods like the Begg and Egger tests that are not developed for DTA meta-analyses. If review authors wants to use statistical methods for investigation publication bias, Deeks' test is recommended for DTA meta-analyses and should be preferred (Van Enst, 2014). In addition, downgrading can be considered when published evidence is limited to few small studies, in particular, if they support a presumed hypothesis and were funded by a body with a vested interest in a particular diagnostic method (Brozek 2009).

5.3. Quality of the evidence of direct benefits, adverse effects or burden of the test

Author(s): Miranda Langendam

Mariska Tuut

Here the GRADE for management decisions approach (KCE process book) can be used (similar to evaluating adverse effects of treatment interventions).

In GRADEpro GDT this is called: *What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?*

5.4. Quality of the evidence of the natural course of the condition and the effects of clinical management guided by the test results

Author(s): Miranda Langendam

Mariska Tuut

For rating the certainty of the evidence of the natural course the GRADE for prognosis approach can be used (Iorio 2015) and for the effects of clinical management GRADE for management decisions (see KCE process book [13]). The evidence about clinical management may be retrieved from other clinical questions in the guideline.

In GRADEpro GDT this is called: *What is the overall certainty of the evidence of effects of the management that is guided by the test results?*

Example: In the WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention, cryotherapy was one of the treatment options after a positive screening test. Applying GRADE for interventions, the certainty in the estimates for the effects of cryotherapy on patient important outcomes (e.g. mortality) - coming from observational studies with high risk of bias - was very low (WHO 2013 – supplemental material)

[13] http://processbook.kce.fgov.be/node/118

5.5. Certainty of the link between test results and management decisions

Author(s): Miranda Langendam

Mariska Tuut

No specific guidance provided. It is important to express how certain the guideline panel is that the result will be followed by clinical management or other action, and if this certainty is based on evaluation of the evidence or assumptions.

In GRADEpro GDT this is called: How certain is the link between test results and management decisions?

Example: How certain are we that a diagnosis of chronic kidney disease is followed by appropriate cardiovascular risk management to prevent cardiovascular events, kidney failure, and mortality?

5.6. Overall quality of evidence

Author(s): Miranda Langendam

Mariska Tuut

The overall rating of the certainty of the evidence about the effects of testing and subsequent management decisions on patient-important outcomes should be based on the certainty of the evidence for the weakest link in the chain of evidence used to estimate those effects.

In GRADEpro GDT this is called: *What is the overall certainty of the evidence of effects of the test?*

Example: In the WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention the authors had very low certainty in some of the steps that were part of the linked bodies of evidence, therefore they rated the overall certainty as very low (WHO 2013, Santesso 2016)

6. Recommendations

Author(s): Miranda Langendam

Mariska Tuut

Going from evidence to recommendation requires judgment of the following criteria (included in GDT): priority of the problem, test accuracy, desirable effects, undesirable effects, certainty of the evidence of test accuracy, direct effects, clinical management effectiveness and link between results and management, overall certainty of effects, values, balance of effects, resources required, certainty of evidence of required resources, cost-effectiveness, equity, acceptability and feasibility. Ideally, the judgments for these criteria are supported by research evidence (systematic reviews). If this is considered too resources intensive, this should be stated.

6.1. Four key factors influence the strength of a recommendation

Author(s): Miranda Langendam

Mariska Tuut

The four key factors influence the strength of a recommendation are:

a. Values

In GRADEpro GDT this is called: *Is there important uncertainty about or variability in how much people value the main outcomes?*

The greater the variability in values and preferences, or uncertainty in values and preferences, the more likely a weak recommendation is warranted.

In the context of tests, how patients value the main patient important outcomes includes adverse effects and burden associated with the test itself, as well as the downstream outcomes of the linked treatment interventions. For example, some patients are highly averse to blood sampling and intravenous lines and others do not mind as much. Patients who fear closed spaces will value not having an MRI and might prefer a CT scan in an open CT. Examples of undesirable effects of downstream consequences are adverse effects or complications of the treatment following a positive test result.

b. Balance between the desirable and undesirable effects

In GRADEpro GDT this is called: *Does the balance between desirable and undesirable effects favor the intervention or the comparison?*

The larger the difference between the desirable and undesirable consequences, the more likely a strong recommendation is warranted. The smaller the net benefit and the lower the certainty for that benefit, the more likely a weak recommendation is warranted.

For tests this judgment is based on the results of either formal or informal modeling of the anticipated desirable and undesirable effects of the test on the patient important outcomes. This includes outcomes related to the burden or direct positive effects of the test.

c. Resource use

In GRADEpro GDT this is called:

How large are the resource requirements (costs)?

What is the certainty of the evidence of resource requirements (costs)?

Does the cost-effectiveness of the intervention favor the intervention or the comparison?

The higher the costs of an intervention – that is, the more resources are consumed – the less likely a strong recommendation is warranted.

For tests, judgments about resource use are the same as for other interventions.

d. Equity, acceptability and feasibility

In GRADEpro GDT this is called:

What would be the impact on health equity?

Is the intervention acceptable to key stakeholders?

Is the intervention feasible to implement?

For tests, assessments of equity, acceptability and feasibility include consideration of both the test and linked treatment interventions.

7. Sources/references

- Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844–1847.
- Bossuyt PM, Irwig L, Craig J, Glasziou P: Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006, 332:1089-1092. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1458557/pdf/bmj33201089.pdf
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015:151516. doi: 10.1148/radiol.2015151516. PubMed PMID: 26509226. http://bmjopen.bmj.com/content/6/11/e012799.long
- Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, Helfand M, Ueffing E, Alonso-Coello P, Meerpohl J, Phillips B, Horvath AR, Bousquet J, Guyatt GH, Schünemann HJ; GRADE Working Group. Grading quality of evidence and strength of recommendations in clinical practice

guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009 Aug;64(8):1109-16. doi: 10.1111/j.1398-9995.2009.02083.x. Epub 2009 May 29. http://onlinelibrary.wiley.com/doi/10.1111/j.1398-9995.2009.02083.x/epdf

- Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;58:882–893

- Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, Schünemann HJ, Bossuyt PM, Leeflang MM, Langendam MW. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014 Jul;67(7):760-8. doi: 10.1016/j.jclinepi.2014.01.006. Epub 2014 Apr 13. http://www.sciencedirect.com/science/article/pii/S0895435614000444?via%3Dihub

- 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. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3126717/pdf/1748-5908-6-62.pdf

- Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, McGinn T, Hayden J, Williams K, Shea B, Wolff R, Kujpers T, Perel P, Vandvik PO, Glasziou P, Schunemann H, Guyatt G. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015 Mar 16;350:h870. doi: 10.1136/bmj.h870. PMID: 25775931

- Korevaar DA, van Es N, Zwinderman AH, Cohen JF, Bossuyt PM. Time to publication among completed diagnostic accuracy studies: associated with reported accuracy estimates. BMC Med Res Methodol. 2016 Jun 6;16:68. doi: 10.1186/s12874-016-0177-4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896017/pdf/12874_2016_Article_177.pdf

- Kapur VK, Auckley DH, Chowdhuri S, Kuhlmann DC, Mehra R, Ramar K, Harrod CG. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep

Medicine clinical practice guideline. J Clin Sleep Med. 2017;13(3):479–504.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5337595/pdf/jcsm.13.3.479.pdf

- Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, Bossuyt P, Garg AX, Lelgemann M, Bühler D, Brozek J, Schünemann HJ. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PLoS One. 2015 Oct 16;10(10):e0134553. doi: 10.1371/journal.pone.0134553. eCollection 2015. PMID:26474310
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4608675/pdf/pone.0134553.pdf

- Nielsen LH, Ortner N, Norgaard BL, Achenbach S, Leipsic J, Abdulla J. The diagnostic accuracy and outcomes after coronary computed tomography angiography vs. conventional functional testing in patients with stable angina pectoris: a systematic review and meta-analysis. European Heart Journal Cardiovascular Imaging 2014;15:961-71.

- David Samson, M.S., Blue Cross and Blue Shield Association, Karen M. Schoelles M.D., S.M., FACP, ECRI Institute Health Technology Assessment Group (Chapter 2). Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.

- 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
Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, Scholten R, Langendam M, Leeflang MM, Akl EA, Singh JA, Meerpohl J, Hultcrantz M, Bossuyt P, Oxman AD; GRADE Working Group. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016 Aug;76:89-98. doi: 10.1016/j.jclinepi.2016.01.032. Epub 2016 Feb 27.

- Schünemann 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;336:1106–1110. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2386626/pdf/bmj-336-7653-analysis-01106.pdf

- van Enst WA, Naaktgeboren CA, Ochodo EA, de Groot JA, Leeflang MM, Reitsma JB, Scholten RJ, Moons KG, Zwinderman AH, Bossuyt PM, Hooft L. Small-study effects and time trends in diagnostic test accuracy meta-analyses: a meta-epidemiological study. Syst Rev. 2015 May 9;4:66. doi: 10.1186/s13643-015-0049-8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450491/pdf/13643_2015_Article_49.pdf

- van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. BMC Med Res Methodol. 2014 May 23;14:70. doi: 10.1186/1471-2288-14-70. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4035673/pdf/1471-2288-14-70.pdf

- Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36. doi: 10.7326/0003-4819-155-8-201110180-00009.

- World Health Organisation. WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. WHO 2013 (available from http://www.who.int/reproductivehealth/publications/cancers/screening_and_treatment_of_precancerous_lesions/en/)