17. Developing guideline recommendations for tests or diagnostic tools
17. Developing guideline recommendations for tests or diagnostic tools

17.1 Introduction

17.2 Evidence pertaining to diagnostic tests

17.2.1 What is a diagnostic accuracy study?

17.2.2 What is a diagnostic test randomised controlled trial (D-RCT)?

17.2.3 Diagnostic accuracy studies or diagnostic randomised controlled trials?

17.3 Formulating diagnostic test questions

17.4 Systematic reviews of diagnostic accuracy studies

17.5 GRADE evidence profiles for diagnostic tests

17.5.1 What is GRADE for diagnostic tests?

17.5.2 GRADE evidence profiles for diagnostic tests

17.6 Evidence-to-decision frameworks for diagnostic tests

17.6.1 What are evidence-to-decision frameworks for diagnostic tests?

17.6.2 Components of evidence-to-decision frameworks for diagnostic tests

17.6.3 When is evidence from test accuracy studies sufficient to develop a recommendation?

17.7 Useful resources

Acknowledgements

References
17. Developing guideline recommendations for tests or diagnostic tools

17.1 Introduction

The prior chapters of the WHO Handbook for Guideline Development (1) have addressed developing recommendations for interventions. While tests can be considered interventions, typically research that concerns tests is in the form of accuracy studies (2,3). Thus, this chapter will address how to develop guideline recommendations for tests from accuracy studies when direct evidence about the test’s effect on patient important outcomes is lacking. This chapter is applicable to screening, monitoring and diagnostic tests, but for clarity we will refer to tests as “diagnostic tests” and test accuracy studies as “diagnostic accuracy studies”.

17.2 Evidence pertaining to diagnostic tests

17.2.1 What is a diagnostic accuracy study?

A diagnostic accuracy study is a study that determines the ability of a particular test (the index test) to correctly classify a patient as having the disease compared with a reference standard test. The reference standard test (sometimes referred to as the gold standard test) should be a test that best defines a person as diseased or not diseased (with ethical and feasibility considerations in mind). In a diagnostic accuracy study, a defined target population undergoes an index test, as well as a reference standard test. The number of true positives, false positives, true negatives, and false negatives is then calculated (Figure 1). From these estimates, the sensitivity and specificity of a test can be calculated, as can other measures of diagnostic accuracy such as positive (PPV) and negative predictive values (NPV) and diagnostic odds ratios. Note that PPV
and NPV, unlike sensitivity and specificity, are affected by the prevalence of a disease.

- “True positives” refers to the number of people that the index test correctly identified as having the condition (i.e. the number of participants that tested positive with both the index test and reference standard).
- “False positives” refers to the number of people that the index test incorrectly identified as having the condition (i.e. the number of people that tested positive with the index test but negative with the reference standard).
- “True negatives” refers to the number of people that the index test correctly identified as not having the condition (a negative result with both the index and the reference standard test).
- “False negatives” refers to the number of people that the index test incorrectly identified as not having the condition (a negative result on the index test but positive with the reference standard).

It should be noted that unlike a randomized, controlled trial (RCT), diagnostic accuracy studies contain only one participant group and are not randomised: all participants undergo both the index and the reference standard test (note that some diagnostic accuracy studies do not subject participants to both the index and reference standard tests, but they are subject to verification bias (4)). Often, however, the participant group may be further stratified into subgroups (e.g. children, HIV+ patients) which is useful to assess sensitivity and specificity of a test in specific populations.
17.2.2 What is a diagnostic test randomised controlled trial (D-RCT)?

Diagnostic tests can also be assessed as interventions in RCTs where participants are randomised to receive or not to receive a test. Thus, as in all RCTs, there are two or more groups and patient-important outcomes such as mortality and morbidity are assessed. In this way, the outcomes generated for D-RCTs differ from the outcomes generated in diagnostic accuracy studies (sensitivity and specificity). D-RCTs can randomise participants to a test, such as randomising participants with low back pain to an x-ray or no x-ray (5). Alternatively, D-RCTs can randomise participants to a test-treatment strategy, such as randomising participants with heart failure to B type natriuretic peptide (BNP) guided management or to usual management (6). Further details of D-RCTs are beyond the scope of this chapter; more information is available in other sources (7, 8).

17.2.3 Diagnostic accuracy studies or diagnostic randomised controlled trials?

Diagnostic accuracy studies generate measures of the accuracy of a test to diagnose a target disease such as sensitivity and specificity, whereas D-RCTs assess the effectiveness of a test on patient-important outcomes. So what evidence should we use as the basis for guideline recommendations? Ideally, diagnostic test recommendations should be based on D-RCTs as they are the optimal way to assess a diagnostic strategy (9). When such studies are available, evidence synthesis techniques and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system for interventions should be used to formulate recommendations (see Chapters 7 to 10 of the WHO Handbook for Guideline Development (1)). Unfortunately, D-RCTs are rare, whereas diagnostic accuracy studies are common (2,3). Thus, this chapter focuses on the more common situation of basing guideline recommendations primarily on diagnostic accuracy studies. It should also be noted that even when D-RCTs are available, diagnostic accuracy studies are often useful for developing recommendations and thus, when available they should also be identified and their data synthesised. The use of both D-RCT data and data from diagnostic accuracy studies is advantageous as health professionals often want information on the accuracy of a test, and the ability to accurately define disease is an important part of many treatment guidelines.
17.3 Formulating diagnostic test questions

The first step in the production of a guideline recommendation is to define the question to be addressed by the recommendation. This question should reflect clinical or public health uncertainty. When one develops a recommendation for diagnostic tests, both a PICO (Population, Intervention, Comparator, Outcome) and a PIRT question (Participants, Index Test(s), Reference Standard, Target condition) should be developed. Structuring questions in the format PICO or PIRT forms the basis for search strategies for D-RCT and for diagnostic test accuracy studies, respectively. Furthermore, at this stage, guideline developers should classify their outcomes as either “critical” or “important but not critical” for their guideline. For guidelines on diagnostic tests, guideline developers should consider accuracy (sensitivity and specificity), as well as the potential side effects of tests and other relevant considerations (e.g. time to conduct the test, time to get results, cost of the test). Note that guideline developers should also consider patient-important outcomes such as mortality: these outcomes are typically determined from intervention studies (D-RCT and observational studies), the methods for which are covered elsewhere in the WHO handbook for Guideline Development (1).

A PICO question should always be formulated because ideally diagnostic test recommendations should be based on D-RCT evidence comparing the effect of a test as an intervention and this helps focus the recommendation on patient-important outcomes rather than only on test accuracy. For instance, a test may be accurate but too invasive, too costly or take too long to generate results to recommend it. Frequently, however, there is no D-RCT evidence to support test recommendations and thus diagnostic accuracy studies are used. This chapter focuses on the production of guideline recommendations from diagnostic accuracy studies and will thus focus on PIRT and developing guideline recommendations based exclusively on diagnostic accuracy studies.

An example of a diagnostic test question in PICO format is: “In patients with smear-positive TB, does the use of line probe assays to diagnose drug resistance lead to lower mortality compared with conventional culture-based drug-susceptibility testing?”. The corresponding PICO is: P, patients with smear-positive TB; I, line probe assays; C, culture-based drug-susceptibility testing; and O, mortality.
An example of a diagnostic accuracy question in PIRT format is “In patients with smear-positive tuberculosis, are line probe assays better than culture-based drug susceptibility tests at diagnosing multi-drug resistance TB (MDR-TB)?”. The corresponding PIRT is: P, patients with smear positive TB; I, line probe assays; R, culture-based drug-susceptibility testing; and T, MDR-TB.

17.4 Systematic reviews of diagnostic accuracy studies

Upon formulation of PICO and PIRT questions, a systematic review is performed to identify, appraise and (if appropriate), produce a meta-analysis of sensitivity and of specificity. The steps to complete a systematic review of diagnostic accuracy studies follow the same structure as that used to complete a systematic review of intervention studies. Reporting guidelines for systematic reviews are available (10), as is a structured tool to assess the quality of a systematic review (ROBIS) (11), and these are applicable to reviews of diagnostic accuracy studies. Note that the AMSTAR-2 tool used to assess the quality of systematic reviews of interventions is not recommended for diagnostic accuracy systematic reviews (12).

There are important considerations when conducting a systematic bibliographic database search for diagnostic accuracy studies. Firstly, search filters that are often applied to systematic searches are not advised as they can miss a considerable number of relevant studies (13). Other important considerations include consulting with clinical and methodological topic experts to generate key search terms, involving a medical librarian, identifying PubMed MeSH terms, and comparing one’s draft search strategy to a search strategy used in published systematic reviews. Ideally, no language restrictions should be placed on the search; unpublished (including trial registers) and grey literature, ongoing studies, as well as conference abstracts should be sought; and authors of primary studies should be contacted if questions or concerns arise in data extraction or assessment. As for any systematic review, studies and data should only be included if there is sufficient information to permit assessment of their risk of bias. Importantly, systematic reviews of diagnostic test accuracy should follow the PRISMA checklist for reporting diagnostic test accuracy reviews (14).
17.5 GRADE evidence profiles for diagnostic tests

17.5.1 What is GRADE for diagnostic tests?

GRADE was originally developed to assess evidence pertaining to interventions, however subsequently the GRADE Working Group developed a system for assessing the quality of evidence pertaining to diagnostic test accuracy (9). Broadly, the GRADE approach is a structured way to assess the quality (certainty) of a body of evidence to address a specific question. In the GRADE approach for interventions, RCTs start as high-quality evidence, whereas observational studies start as low-quality evidence (15); the quality of evidence is then downgraded or upgraded according to specific criteria (see Chapter 9 of the WHO handbook for guideline development (1)). In the GRADE approach for diagnostic accuracy studies, the body of evidence starts as high quality (certainty), although such studies provide only indirect evidence for patient-important outcomes (9). This is different from GRADE for interventions where observational studies start as low-quality evidence. This chapter will outline guidance for WHO staff using the GRADE system for assessing the quality of evidence pertaining to diagnostic test accuracy.

17.5.2 GRADE evidence profiles for diagnostic tests

GRADE evidence profiles are structured, domain-based tables that outline factors that can affect the quality (certainty) of a body of evidence. Importantly, they are outcome-based: they incorporate all the available evidence for specific outcomes. In the case of diagnostic accuracy studies, these outcomes are sensitivity (true positives and false negatives) and specificity (true negatives and false positives). GRADE evidence profiles for diagnostic accuracy follow the same structure as evidence profiles for interventions: they contain summaries of the diagnostic accuracy estimates and the number of studies and patients contributing to these estimates. They also include five domains for consideration of the certainty of the evidence for each outcome: risk of bias, indirectness, inconsistency, imprecision and publication bias. All the domains except publication bias are assessed individually to determine if very serious, serious or no serious limitations exist. Publication bias
is considered either “not detected” or “highly suspected”. If there are no 
serious limitations across any of the five domains, the quality of evidence is 
not downgraded. If serious limitations exist in one of the five domains, the 
evidence is downgraded by one level, from high to moderate. If very serious 
limitations exist for one domain or serious limitations are present in two 
domains, the evidence is downgraded by two levels, from high to low quality. 
Similarly, if three domains contain serious limitations (or one domain with 
very serious limitations and another with serious) the evidence is down-
graded by three, to very low.

Although the five domains that decrease the certainty of the effect esti-
mates of test accuracy studies are the same as those that affect the certainty 
of intervention studies, they require different operationalization. Further, 
the GRADE evidence profiles for diagnostic accuracy have two groups of 
outcomes: true positives and false negatives (sensitivity) and true negatives 
and false positives (specificity) (Figure 2). The following sections will outline 
how to assess the five domains as they pertain to diagnostic test studies.

17.5.2.1 Risk of bias
To assess the risk of bias (also called study limitations) of individual diag-
nostic accuracy studies, the QUADAS-2 (Quality Assessment of Diagnos-
tic Accuracy Studies) tool (17) should be used. This should be completed 
by the systematic review authors. This tool assesses the risk of bias of 
individual studies across four domains: participant selection, index test, 
reference standard, and flow and timing across both sensitivity (true posi-
tives and false negatives) and specificity (true negatives and false posi-
tives). The full QUADAS-2 tool also contains an applicability domain; 
however, this domain should not be used to downgrade evidence when 
assessing risk of bias as this is covered in the indirectness domain of the 
GRADE system when the certainty of the body of evidence is assessed. 
The QUADAS-2 tool contains a series of signalling questions for each 
domain, which the appraisers use to judge each domain as high, low or 
unclear risk of bias.

Results from the risk of bias assessment (of each included study) with 
the QUADAS-2 tool can be displayed graphically (Figure 3) (16), and used 
to assess the overall risk of bias for a body of evidence concerning an 
outcome.
Figure 2. Example of a GRADE evidence profile for diagnostic tests

Accuracy of LPAs for diagnosing MDR-TB on all specimen types by direct and indirect testing

**Participants:** Patients with signs and symptoms of TB

**Prior testing:** None

**Role:** Replacement test for culture-based drug susceptibility testing

**Settings:** Intermediate or specialized laboratories

**Reference standard:** Culture based drug susceptibility testing

**Studies:** Case-control or cohort studies comparing LPAs with a reference standard

**Index [new] tests:** GenoType MTBDRplus version 1 assay (Hain Lifesciences, Nehren, Germany); GenoType MTBDRplus version 2 assay (Hain Lifesciences, Nehren, Germany); Nippon NTM MDR-TB detection kit 2 (Nippon, Tokyo, Japan). The tests were performed on all types of specimens using direct and indirect testing.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Number of studies (number of patients)</th>
<th>Study design</th>
<th>Factors that may decrease the quality of evidence</th>
<th>Effect per 1 000 patients tested (number of patients)</th>
<th>Test accuracy quality of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives (patients with MDR-TB)</td>
<td>60 studies (4 248 patients)</td>
<td>Cohort and case-control type studies*</td>
<td>Serious* Not serious* Not serious* Not serious* Not serious* Not serious* Not serious*</td>
<td>9 (9-9) 45 (45-47) 93 (90-96)</td>
<td>Moderate</td>
</tr>
<tr>
<td>False negatives (patients incorrectly classified as not having MDR-TB)</td>
<td>1 (1-1)</td>
<td>1 (1-1)</td>
<td>4 (3-5)</td>
<td>7 (5-10)</td>
<td>Low</td>
</tr>
<tr>
<td>True negatives (patients without MDR-TB)</td>
<td>60 studies (8 785 patients)</td>
<td>Cohort and case-control type studies*</td>
<td>Serious* Not serious* Not serious* Not serious* Not serious* Not serious* Not serious*</td>
<td>983 (977-986) 943 (938-949) 994 (988-996)</td>
<td>Low</td>
</tr>
<tr>
<td>False positives (patients incorrectly classified as having MDR-TB)</td>
<td>7 (4-13)</td>
<td>7 (4-12)</td>
<td>6 (4-12)</td>
<td>Moderate</td>
<td></td>
</tr>
</tbody>
</table>

*For interpretation of the numbers, see the manual for GRADE (15).
There are no specific rules for what proportion of low, high or unclear risk of bias assessments with the QUADAS-2 tool constitutes very serious, serious or no serious limitations for the GRADE assessment of the certainty of the body of evidence for each outcome. This is a judgement that needs to be made in the context of individual situations, however the reasons for the decision need to be explicit, transparent and clear. In Figure 3, for example, the body of evidence was assessed as having serious limitations because: “The sampling of patients (participants) was often not stated”, reflecting the unclear risk of bias in the majority of included studies (approximately 60%) in the patient selection domain of the QUADAS-2 tool (16). In addition, “it was often not stated whether investigators were blinded between the index and reference standard” (unclear risk of bias in the majority of included studies (approximately 70%) in the index test and reference standard domains of the QUADAS-2 tool) (16).

**Figure 3. Example of a display of risk of bias assessments across studies**


### 17.5.2.2 Indirectness

The assessment of the indirectness of a body of evidence refers to two concepts.

1. How well does the underlying body of evidence match the clinical or public health question (PIRT)?
2. How certain can the reviewer be that the consequences of using the test will lead to improvement in patient-important outcomes?

These two questions should both be considered in the assessment of indirectness; one or both can contribute to downgrading the evidence for
indirectness. For instance, if the underlying body of evidence does not match the clinical or public health PIRT, the evidence should be downgraded, even if there are no concerns regarding question 2 (and vice versa). To address the first question, the Guideline Development Group needs to compare the PIRT question they generated in Section 17.3 to the PIRT of the underlying body of evidence (from their systematic search). Differences in the participants, index and/or reference standard tests and setting between the PIRT developed a priori by the guideline developers and the PIRT of the underlying evidence may mean that the results from the underlying evidence do not translate to the intended patients or populations. For instance, consider if a Guideline Development Group commissioned a systematic review to determine the accuracy of brain natriuretic peptide (BNP) to diagnose heart failure in patients with signs and symptoms of heart failure presenting to general practice, compared to a reference standard of echocardiography. The Guideline Development Group may consider downgrading if the available evidence only included patients with shortness of breath and no other symptoms of heart failure, did not use echocardiography as a reference standard (be careful not to double count this in the risk of bias assessment), or included patients presenting to the emergency department. Evidence may be additionally downgraded if it was conducted in a high-resource setting and the guideline is intended for low-resource settings.

It should be noted that PIRT questions that include two or more index tests present an exceptional situation. In these circumstances, an additional concept pertaining to indirectness should be considered: were the two index tests compared directly in the included studies using the same reference standard? If this is not the case, downgrading should occur.

Furthermore, the outcomes from diagnostic accuracy studies, sensitivity and specificity, are inherently indirect: improved diagnostic accuracy does not always translate into improvements for patients or populations (2). Improved diagnostic accuracy can lead to benefits for patients but relies on the following assumptions:

- A test with an increased number of true positives (TP) will benefit patients if effective treatment for the disease is available, the benefits of treatment outweigh the harms and all patients receive the treatment. Also, benefit will only be seen if effective health services exist which are able to deliver the treatment to patients.
- A test with fewer false negatives (FN) will benefit patients by minimising delayed diagnoses if the natural history of the disease leads to morbidity and mortality.
A test with an increased number of true negatives (TN) may benefit patients by providing reassurance and sparing from unnecessary treatment. However, patients are not uniformly reassured by true negative test results: they may seek another diagnosis to explain their symptoms.

A test with fewer false positive (FP) results will benefit patients if treatment and/or additional testing lead to adverse effects and/or the diagnostic label leads to patient anxiety.

To assess indirectness, the Guideline Development Group needs to determine how confident they are that the above assumptions are true, ideally with evidence to support their decisions. For instance, evidence to address the TP and FP assumption should come from RCTs of a treatment (and its adverse effects) of the target disease. Evidence to address the FN assumption can come from prognostic (natural history) studies and/or the control arms of RCTs.

When assessing the indirectness of diagnostic accuracy evidence using the GRADE system, guideline panels need to determine if very serious, serious or no serious indirectness exists by answering the two questions stated at the beginning of this section. Indirectness with respect to either or both of these questions can lead to downgrading the evidence. As is the case for all components of the GRADE system for diagnostic accuracy studies, indirectness needs to be determined for both sensitivity and specificity. In reality, the first indirectness question (how well does the underlying body of evidence match the PICO question?) is difficult to apply to sensitivity and specificity separately, and thus sensitivity and specificity are often downgraded (or not downgraded) together. However, this is not the case when addressing the second indirectness question where sensitivity and specificity are often downgraded independently of each other. For instance, if an effective treatment is not available, a test that produces more true positive results will not necessarily translate to improved patient outcomes. This situation would lead to downgrading sensitivity (true positives and false negatives) but would not necessarily lead to downgrading specificity as the latter is dependent on the assumptions (as stated above) surrounding the number of true negatives and false positives.

17.5.2.3 Inconsistency
This domain in the GRADE approach refers to an assessment of the inconsistency of effect estimates (in our case sensitivity and specificity) across included studies. Substantial heterogeneity among different test accuracy
Studies is common, expected, and generally not considered to be due to chance alone (19). There are a number of factors that contribute to heterogeneity in meta-analyses of diagnostic accuracy studies: chiefly, the variation in the threshold used to define a participant as having or not having the disease (often referred to as the “positive threshold”) (20). Given the expected heterogeneity, meta-analyses of diagnostic accuracy should be performed with a random-effects model rather than a fixed-effects model (19). For similar reasons, conventional tests investigating heterogeneity such as the I² test (21), are infrequently used and are not endorsed by Cochrane (19).

Lastly, as is the case for all five components of the GRADE assessment of the certainty of evidence, sensitivity and specificity should be considered and assessed separately.

For the investigation of inconsistency in diagnostic accuracy meta-analyses, Guideline Development Groups should address the following questions.

1. How similar are the point estimates across the primary studies?
2. Do the confidence intervals overlap?

To address these questions Guideline Development Groups can use forest plots. Consider the forest plots in Figure 4. The sensitivity point estimates are all above 72% with overlapping confidence intervals, apart from one outlying study. Conversely, the point estimates for specificity are scattered between 12% and 88%. In this example, specificity should be downgraded for inconsistency. For sensitivity, if there is an apparent reason for the outlying study (e.g. poor quality as assessed with QUADAS-2, use of a different positive threshold, smaller study, different PIRT or study setting) then sensitivity need not be downgraded.

Importantly, inconsistency should only lead to downgrading if it is unexplained. If inconsistency can be explained, for instance by factors explored with sensitivity analyses or meta-regression, then it should not lead to downgrading.

Lastly, it is important to note that differences in PIRT between the systematic review and the guideline outcome should be addressed in the indirectness domain of GRADE and not as inconsistency.

17.5.2.4 Imprecision
This component guides panels to judge whether the pooled sensitivity and specificity are precise enough to support a recommendation. This assessment
Figure 4. Example of a forest plot of sensitivity and specificity

Reproduced from Mustafa 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 Mar;132(3):259-65. doi: 10.1016/j.ijgo.2015.07.024. Epub 2015 Nov 12. (22). Licences under the Creative Commons Attribution, Non-Commercial, No Derivative works licence (CC BY-NC-ND 4.0); http://creativecommons.org/licenses/by-nc-nd/4.0/). Note that the I^2 and Q estimates have been removed from the original figure. Pooled estimates are derived from a hierarchical model.
concerns the width of confidence intervals surrounding the pooled sensitivity and the pooled specificity. Guideline Development Groups can use the following two concepts to guide their judgement on the width of confidence intervals:

1. specify a priori a width of the confidence interval that constitutes imprecision; or
2. consider how a recommendation may change if the sensitivity and specificity were at the upper or lower limit of their respective confidence intervals.

The assessment of the imprecision of a test’s sensitivity and specificity is a judgement that may be influenced by a number of factors. Most significantly, guideline panels should consider the role of the test and the prevalence of the disease. To produce more transparent guidelines, some previous WHO Guideline Development Groups have used explicit thresholds for assessments of imprecision. For instance, a guideline panel addressing a question of the accuracy of line probe assays for detecting rifampicin resistance in patients with signs and symptoms of TB considered estimates to be imprecise if the confidence intervals of the pooled sensitivity were wider than 10% in either direction (16). On the other hand, they considered the pooled specificity to be imprecise if confidence intervals were wider than 5% in either direction (16).

If a threshold is going to be used it should be specified a priori and be generated in line with the role of the specific test. For instance, different thresholds may be selected for tests that are used to rule out disease (where a high level and precise estimate of sensitivity are desirable); conversely tests used to rule in disease rely on a high level of specificity. Ideally, any threshold that is used should be informed by evidence, particularly regarding the consequences of varying test results on patient-important outcomes. For example, modelling of the number of false positives and follow up consequences could be used if specificity is lower than a chosen threshold. Furthermore, there are some unique situations where wide confidence intervals around measures of diagnostic accuracy do not necessarily lead to a downgrading of the certainty of the evidence. For instance, in a condition that has a low prevalence, wide confidence intervals around sensitivity may not lead to real concern with imprecision for TP and FN.

Thresholds need not be applied for all guidelines. A Guideline Development Group may also consider how their recommendations may change if they assumed the sensitivity (or specificity) was the lower limit of the con-
confidence interval. For instance, consider a pooled sensitivity of 90% with a confidence interval of 60% to 98%; although a sensitivity of 90% seems high, if the sensitivity was 60% then guideline developers may not consider this an adequate sensitivity to support the use of a test. If guideline panels choose to assess imprecision this way, it is desirable for the change in TP, FP, TN and FN to be determined, and, importantly, the consequences of these changes modelled: for example, the number of people successfully treated at the lower limit of the confidence interval.

17.5.2.5 Publication bias

The final component of the assessment of certainty of the evidence for diagnostic tests is publication bias. This refers to the increased likelihood that studies with statistically significant or favourable results will be published compared to results with non-significant or unfavourable results (23). This can, in turn, lead to syntheses of incomplete sets of the evidence and produce summary results potentially biased towards favourable treatment effects (23). A substantial amount of research exists on publication bias in systematic reviews of RCTs, however little research exists for systematic reviews of diagnostic accuracy (19).

Currently there is no widely agreed upon statistical test to help identify publication bias in diagnostic accuracy studies and previous WHO guidelines have not assessed it (24,25). Guideline developers should ensure that searches for evidence are comprehensive (including trial registers and grey literature), acknowledge any limitations in searching, and report any evidence to suggest that important information may not have been retrieved.

17.5.2.6 Assessing the certainty of the body of evidence across outcomes

After completing the assessment of the five domains for each critical outcome (for PICO questions) and for sensitivity and specificity (for PIRT questions), the overall assessment of the certainty of the body of evidence can be determined for each of the PICO and PIRT questions, across outcomes. As is the case for GRADE for interventions, the overall certainty can be assessed as high, moderate, low or very low. Evidence starts as high quality but is downgraded by one (e.g. high to moderate) or two (moderate to very low) levels depending on an aggregate assessment across outcomes. Examples of diagnostic test evidence profiles are available at (http://www.who.int/tb/areas-of-work/laboratory/fllpa_online_annexes.pdf?ua=1) and the GRADE Database of Evidence Tables (https://dbep.gradepro.org/search).
Guideline developers can use GRADEpro software (available at https://gradepro.org/) to present their evidence profiles. This software allows both the presentation of each of the five domains for assessing certainty of evidence and summary of findings (SOF) tables. GRADEpro also allows guideline developers to view the data in many different forms and to assess multiple outcomes including sensitivity, specificity, positive and negative predictive value and other measures of diagnostic accuracy; this facilitates formulation of recommendations.

17.6 Evidence-to-decision frameworks for diagnostic tests

17.6.1 What are evidence-to-decision frameworks for diagnostic tests?

Evidence-to-decision (EtD) frameworks were developed to facilitate the process of formulating recommendations based on various considerations in addition to the benefits and harms of an intervention. These considerations affect the direction (for or against the intervention) and the strength (strong or conditional) of a guideline recommendation. The GRADE system includes explicit guidance on how to use EtD frameworks when developing diagnostic test recommendations (26), in addition to specific guidance for interventions (27) (see Chapter 10 of the WHO handbook on guideline development (1)).

Each of the considerations in the EtD framework should be thoughtfully examined at the beginning of any guideline development process, whether for interventions or for diagnostic tests. Once the general scope and key questions are decided upon, guideline developers need to consider what evidence in addition to diagnostic test accuracy and patient-important health outcomes is needed to enable the Guideline Development Group to formulate a given recommendation. Additional evidence on, for example, feasibility or acceptability, may be needed to inform the recommendation. Guideline developers must specify the perspective they are taking, whether a population or an individual patient perspective, prior to seeking evidence and populating the EtD framework.
### Table 1. Components of the GRADE evidence-to-decision framework

<table>
<thead>
<tr>
<th>Question</th>
<th>Judgements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the problem a priority</td>
<td>Don't know</td>
</tr>
<tr>
<td>Test Accuracy</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Desirable effects</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Undesirable effects</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Certainty of test accuracy</td>
<td>No studies</td>
</tr>
<tr>
<td>Certainty of critical or important outcomes</td>
<td>No studies</td>
</tr>
<tr>
<td>Certainty of the management guided by test results</td>
<td>No studies</td>
</tr>
<tr>
<td>Certain of link between test results and management</td>
<td>No studies</td>
</tr>
<tr>
<td>Overall certainty about effects of test</td>
<td>No studies</td>
</tr>
<tr>
<td>Value</td>
<td>Important uncertainty or variability</td>
</tr>
<tr>
<td>Balance of desirable and undesirable effects</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Resource requirements</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Certainty of resource requirements?</td>
<td>No included studies</td>
</tr>
<tr>
<td>Cost-effectiveness</td>
<td>No included studies</td>
</tr>
<tr>
<td>Equity</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Acceptability</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Feasibility</td>
<td>Don’t know</td>
</tr>
</tbody>
</table>
17.6.2 Components of evidence-to-decision frameworks for diagnostic tests

The assessments that make up the GRADE EtD framework for diagnostic tests are listed in Table 1 and detailed below (26).

17.6.2.1 Is the problem a priority?
The first component of the EtD framework is a judgement regarding the priority of the problem being addressed in the recommendation question. Here Guideline Development Groups must consider evidence regarding the prevalence, incidence, morbidity, mortality and cost of the disease the recommendation is addressing. The burden of under- or over-diagnosing patients can also be highlighted. This component is not necessary and can be omitted if the guideline or recommendation is not prioritizing treatment or testing options; the problem was clearly a priority or the guideline would not have been undertaken.

17.6.2.2 How accurate is the test?
This component can be answered directly from the pooled sensitivity (TP, FN) and specificity (TN, FP) presented in the GRADE evidence profiles. It should be noted that guideline developers should consider a priori what they will consider “very accurate”, “accurate”, “inaccurate” and “very inaccurate”. Also, they should consider the specific prevalence of the disease that will be considered in this judgement. GRADE evidence profiles allow, and in fact encourage, developers to enter different disease prevalences to see how the number of true positives, false positives, true negatives and false negatives change per 1000 patients (see Figure 2; pre-test probability is synonymous with disease prevalence).

17.6.2.3 How substantial are the desirable and undesirable anticipated effects?
This component asks panels to judge the anticipated benefits and harms from the test in question, including direct effects from the test (e.g. benefits such as faster diagnosis, and harms such as adverse effects from administration of the test). In addition, the possible subsequent effects of the test must be included, for instance effects of treatment after a positive diagnosis and the effect of no treatment or further testing after a negative test. Evidence should inform these downstream effects after a diagnosis, ideally from systematic reviews of D-RCTs. For instance, an EtD table completed to determine if Human Papilloma Virus (HPV) tests should be used to screen for cervical
intra-epithelial neoplasia 2 (CIN2) (a cervical cancer precursor), included evidence from systematic reviews of benefits (e.g. decreased mortality) and harms (e.g. infection, bleeding) from the three possible treatments for CIN2 (22). If systematic review evidence is not available, the potential subsequent effects can be modelled. For instance, the prevalence of the disease in question combined with the sensitivity and specificity can be used to estimate the number of FPs and FNs in a population and to estimate PPV and NPV. These data can assist guideline developers to make a judgement about the undesirable effects of the test.

17.6.2.4 **What is the overall certainty of the test accuracy evidence?**
This component is the judgement of the quality (certainty) of the evidence for diagnostic test accuracy. For this domain of the E2D framework, sensitivity and specificity should be considered as one collective measure of test accuracy.

17.6.2.5 **What is the certainty of the evidence for any critical or important outcomes?**
Critical or important outcomes in the GRADE approach are those related to harms and benefits caused to patients (26) and these are specified a priori by guideline developers (26,28). This component prompts guideline developers to assess the quality (certainty) of evidence on the direct benefits and harms of the test (part of the assessment in section 17.6.2.3). Direct benefits can include faster diagnosis, for example, whereas direct harms refer to specific harms from the test, for instance allergic reactions to radioactive contrast dye. Often these outcomes data will be found in the diagnostic accuracy studies included in the systematic review, however, they may also come from other primary studies or systematic reviews. Panels will have to make a judgement on the quality of evidence from these additional studies (very low, low, moderate or high).

17.6.2.6 **What is the overall certainty of the evidence of effects of the management that is guided by the test results?**
Guideline Development Groups are asked to judge the certainty of evidence for two issues.

1. The evidence supporting treatment and management after a positive diagnosis: a) the quality of evidence supporting the treatment of the target disease (this is particularly relevant for persons who are classified as TP); and b) the quality of evidence for adverse effects of
treatment (this is particularly relevant for persons that test FP). This evidence should be addressed in a systematic review of RCTs or observational studies, with a corresponding GRADE evidence profile for interventions.

2. The evidence supporting the natural history (prognosis) of the target condition: Improvement or deterioration without treatment or further management is relevant to those that test FN. The evidence on the natural history of the target condition should generally come from the control arms of RCTs or observational studies and the quality of evidence is judged using the GRADE approach for questions of prognosis (29).

17.6.2.7 How certain is the link between test results and management decisions?

Guideline Development Groups must make a judgement about the likelihood that the appropriate management (such as treatment decisions) will follow on from test results. Important features of a test, such as test turnaround time and interpretability of results can pose barriers to patients receiving the appropriate treatment after obtaining a test result. Further, there may be factors external to the test that reduce the likelihood of patients receiving appropriate management after a test such as out-of-pocket expenses; access to quality, coordinated services; health literacy; among many others. Guideline Development Groups should consider the literature broadly and seek and include relevant contextual knowledge about the target healthcare settings for potential barriers that prevent appropriate follow up and management after a test. Although it is highly preferable to consider published research studies to inform this judgement, often guideline developers have to rely on their own experience concerning the likelihood a test result is managed appropriately and this can still be considered “high” certainty. For instance, the 2018 American Society of Haematology Guideline for management of venous thromboembolism: diagnosis of venous thromboembolism (30) considered that “with [a pulmonary embolus] (PE) diagnosis, positive results will be treated with anticoagulation (regardless of the chances of false positives)” (30). They assumed this because the “intervention is relatively simple to apply and few patients would be missed in health care systems that are equipped to offer testing” (26,30). The guideline panel thus considered the certainty of the link between the diagnostic test results and management decisions as high.
Chapter 17. Developing guideline recommendations for tests or diagnostic tools

17.6.2.8 What is the overall certainty of the evidence about the effects of the test?
This component requires Guideline Development Groups to make an overall judgement about the certainty (quality) of the evidence, considering the lowest quality of the previous four sections (17.6.2.4 to 17.6.2.7) which reflect the entire test-to-treatment pathway: from the accuracy of the test, to the likelihood patients that test positive get treated, to the effectiveness of treatment.

17.6.2.9 Is there important uncertainty about or variability in how much people value diagnostic accuracy of this test, and the other outcomes associated with the test-to-treat pathway?
This section addresses how much people value the outcomes and how this may affect recommendations. The outcomes of interest include accuracy (sensitivity and specificity) of a test, but also outcomes related to direct benefits and harms from tests and subsequent management of the disease or condition. For instance, Guideline Development Groups should use evidence from persons effected (or potentially effected) by the target condition and its associated test to assign a value to usability and resource requirements of a test, time-to-result, benefits and harms from treatment following a positive test, as well as sensitivity and specificity of the test. As stated in section 17.3 each outcome should be rated as critical or important according to GRADE guidance (26), informed by qualitative studies reflecting patient, healthcare provider and other stakeholders’ perspectives. This will help to guide how a test (or a collection of test strategies) should be used in practice. Patient values may help lead guideline developers to recommend the test to rule out disease and thus prioritize a highly sensitive test. For example, the 2018 American Society of Haematology guideline prioritized a sensitive initial test because of patient desire to rule out a pulmonary embolism (30). Guideline Development Groups should reflect any uncertainty or variability in how patients and healthcare professionals value the outcomes. The Guideline Development Group will often survey its members to gather their views on the relative value of each outcome (although ideally persons effected by the recommendation will be surveyed). For an example, see https://guidelines.gradepro.org/profile/e9600faf-99bc-4ade-9f2f-70bf6e078f9e (in the evidence-to-decision framework tab).
17.6.2.10 **Does the balance between desirable and undesirable effects favour the index test or the comparison?**

Guideline Development Groups are prompted to make an overall judgement about the benefits and harms of the test. This assessment is based on the accuracy of the test, direct benefits (e.g. faster diagnosis) and harms (e.g. radiation) from the test, the benefits and harms from management following the test results, and the certainty of the bodies of evidence informing these assessments. For instance, a WHO guideline on the use of line probe assays to diagnosis multi-drug resistant TB judged the balance of benefits and harms to favour the line probe assay because of its high sensitivity and specificity and subsequent small numbers of FN and FP results, coupled with documented reductions in diagnostic and treatment delays (16). To aid decision-making, modelling can be performed to determine the number of TP, FP, FN and TN based on different disease prevalences (pre-test probabilities). Similarly, the number or rate of harmful effects of tests can be determined and compared and assessed. In the absence of data to support a decision, an assessment should usually be made by the Guideline Development Group. In this situation an assessment of “probably favours” either the intervention or the comparison should be made.

17.6.2.11 **Resource requirements**

Resource requirements for implementing a new test should be considered before a recommendation is issued. Resource requirements are addressed by three questions.

1. How large are the resource requirements?
2. What is the certainty of the evidence on resource requirements?
3. Does the cost-effectiveness of the intervention favour the intervention or the comparison?

When addressing this component, Guideline Development Groups should consider the resource requirements of both the test and subsequent management (e.g. treatment). Because resource cost and affordability are so contextual, the Guideline Development Group may elect to not consider cost in the recommendation, rather explicitly noting that adoption and adaptation of the guideline to the national or sub-national context will require careful consideration of the cost of the test and subsequent management strategies.
17.6.2.12 **Equity, acceptability and feasibility**
Guideline developers must consider the impact of the index test on health equity, the acceptability of a new test by relevant stakeholders and whether the implementation of the test is feasible. Evidence that can be used to address these issues includes qualitative studies on the perspectives of key stakeholders (e.g. patients, health professionals and programme managers). Often, due to the lack of research evidence addressing these components, Guideline Development Groups rely on the experience with the test in different settings. For instance, the successful use of liquid-based cytology as a method of screening in the Kingdom of Saudi Arabia was used to support an assessment of the feasibility of implementing a new cervical cancer screening method in the WHO guideline on cervical cancer screening (22,31). Equity, acceptability and feasibility should all be judged individually.

17.6.2.13 **Developing a recommendation**
A completed EtD framework assists Guideline Development Groups to generate transparent, valid, trustworthy recommendations. Guideline developers need to consider both the direction of the recommendation (whether the test should be recommended or not) and the strength of the recommendation (strong or conditional). How to formulate recommendations, including when to issue a strong or conditional recommendation, is covered in detail in Chapter 10 of the *WHO handbook on guideline development* (1). There are, however, some issues specific to formulating diagnostic test recommendations.

It is common for a discrepancy to exist between the certainty of the evidence pertaining to test accuracy and the certainty of evidence for patient-important outcomes. This is most commonly due to uncertainty surrounding the link between test and treatments and/or uncertainty surrounding the effect of subsequent management. In these situations, where the certainty surrounding diagnostic accuracy evidence is moderate or high, but the certainty surrounding the evidence of downstream management and/or the link between the test and management is low or very low, this uncertainty should be reflected and in most situations a conditional, rather than strong recommendation will be appropriate (26).

17.6.3 **When is evidence from test accuracy studies sufficient to develop a recommendation?**
A 2017 study reported the results of interviews with 24 international experts in evidence and decisions about healthcare-related tests and diagnostic strat-
egies (32). This study concluded that “test accuracy is rarely, if ever, sufficient to base guideline recommendations” and thus evidence-to-decision frameworks are necessary to help developers to consider important issues beyond test accuracy. Diagnostic test experts did, however, note four potential situations when test accuracy is likely to be sufficient to extrapolate the effects of tests on patient-important outcomes (32):

1. when diagnostic noninferiority is sufficient for a decision;
2. when inferences can be made about the impact on patient-importance outcomes;
3. when the accuracy of one test is equivalent or better than the combined accuracy of two tests; and
4. when the primary goal of the guideline is to establish a diagnosis for a condition or to rule out a condition.

In all of these four scenarios, however, there is an assumed link to patient-important outcomes. As such, even if one of these four situations applies, it is still advisable to consider patient-important outcomes when formulating recommendations.

17.7 Useful resources

The Cochrane handbook for systematic review of diagnostic test accuracy: http://methods.cochrane.org/sdt/handbook-dta-reviews (33)

- The GRADEpro website: https://gradepro.org/
- GRADE database of evidence profiles: https://guidelines.gradepro.org/search

Acknowledgements

This chapter was prepared by Dr Jack W. O’Sullivan and Dr Susan L Norris with peer review by Professor Reem Mustafa, Dr Mariska (MMG) Leeflang and Dr Alexei Korobitsyn.
Chapter 17. Developing guideline recommendations for tests or diagnostic tools

References


Chapter 17. Developing guideline recommendations for tests or diagnostic tools


