Guidance on evidence generation for new tuberculosis diagnostics

Draft for Public Comment

Publication Page Placeholder

Contents

Acknowledgements

Development of this document was led by Samuel Schumacher with support from Mikashmi Kohli, Alexei Korobitsyn, Carl-Michael Nathanson, Nazir Ismail, Patricia Hall-Eidson, and Matteo Zignol under the overall direction of Tereza Kasaeva, Director, Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections.

Expert inputs were also provided over multiple rounds of review by the WHO Steering Committee that included Vineet Bhatia (SEARO), Saskia den Boon (IVD), Soudeh Ehsani (EURO), Dennis Falzon (XXX), Kyung Hyun Oh (WPRO), Jean Iragena (AFRO), Avinash Kanchar (XXX), Cecily Miller (XXX), Ernesto Montoro (PAHO), Irena Prat (PQ), Martin van den Boom (EMRO), and Kerri Viney (XXX).

The WHO is also grateful to the Scientific GEG Development Group for their valuable reviews and expert contributions to the development of this guidance document: Alex Kay (Baylor School of Medicine), Alice Zwerling (University of Ottawa), Anis Karuniawati (University of Indonesia, Jakarta), Augusto Geyer (ANVISA, Brazil), Christopher Coulter (Queensland Mycobacterium Reference Laboratory), Chishala Chabala (University Teaching Hospital, Lusaka, Zambia), Claudia Denkinger (University of Heidelberg), Daniela Maria Cirillo (San Raffaele Scientific Institute), Dick Menzies (McGill University), Emily McLean (University of Sydney), Grant Theron (Stellenbosch University), Hoojoon Sohn (Seoul National University College of Medicine), Holger Schünemann (McMaster University), Jon Deeks (University of Birmingham), Madhukar Pai (McGill University), Mark Nicol (University of Western Australia), Monde Muyoyeta (Centre for Infectious Disease Research in Zambia), Nandita Venkatesan (LiveMint), Nora Engel (Athena Institute), Rachel Byrne (Liverpool School of Tropical Medicine), Sabira Tahseen (National TB Control Programme of Pakistan), Shaheed V. Omar (Centre for Tuberculosis, National Institute for Communicable Diseases), Siva Kumar Shanmugam (Department of Bacteriology; National Institute for Research in Tuberculosis), Tejaswini Dharmapuri V (TAG), Thomas Shinnick (Independent Consultant), Valeria Alcon (Health Canada), Violet Chiota (Aurum Institute), Yanlin Zhao (National Tuberculosis Control and Prevention Center, China), Yemisi Takwoingi (University of Birmingham).

Further, the following experts external to WHO are acknowledged for their expert comments and suggestions: XXX

This product was developed with support from the Gates Foundation.

Abbreviations and acronyms

TBD

Executive summary

Narrative under development, noting that the table below will become part of the executive summary

Document section	Key messages
What evidence is needed (Section 4.2)	**Key message 1 –** Define the value proposition and prepare an analytical framework
Accuracy studies (Section 4.3)	**Key message 2 –** Design studies to minimize risk of bias **Key message 3 –** Align selection criteria with the target population **Key message 4 –** Carefully consider selection of settings for participant enrolment **Key message 5 –** Generate evidence on the index test in its intended setting of use **Key message 6 –** Carefully consider and describe specimen processing and testing **Key message 7 –** Formulate a strategy for extrapolation of evidence to excluded populations, settings or specimen types **Key message 8 –** Ensure sufficient sample size to achieve precise estimates **Key message 9 –** Provide a precise description of how the index test was applied **Key message 10 –** Select an appropriate reference standard **Key message 11 –** Include a comparator in the study **Key message 12 –** Report transparently and provide comprehensive analyses **Key message 13 –** Share individual participant data
Other evidence gathered as part of accuracy studies (Section 4.4)	**Key message 14** – Provide a careful analysis of non-positive non-negative results and an assessment of test robustness **Key message 15** – Measure time to result for the index test and comparator **Key message 16** – Evaluate possible procedural harms or burdens associated with testing **Key message 17** – Consider conducting studies on test-positivity rates (diagnostic yield) if the diagnostic intervention may increase access to testing **Key message 18** – Consider additional design and analytical aspects when evaluating diagnostic strategies comprised of more than one test
Evidence to support linkage across the analytical framework	**Key message 19** – Generate evidence to link diagnostic accuracy data to changes in intermediate and final health outcomes
Evidence on patient-important outcomes (Section 4.4)	**Key message 20 –** Consider conducting diagnostic randomized controlled trials when judging tests' effects on health outcomes based on accuracy may not be reliable **Key message 21** – Generate evidence on the effect of the diagnostic intervention on intermediate outcomes **Key message 22** – Generate evidence on the effect of the diagnostic intervention on final outcomes
Evidence beyond health outcomes (Section 4.6)	**Key message 23 –** Conduct research on values (i.e. the relative importance people place on health outcomes) **Key message 24 –** Gather evidence on the resources required to deliver the diagnostic intervention **Key message 25 –** Carry out cost–effectiveness analyses **Key message 26 –** Investigate the impact on health equity **Key message 27 –** Investigate the acceptability of the test **Key message 28** – Investigate the feasibility of implementing the diagnostic intervention

Structure of the document

Section 1 provides the background for the document and defines its purpose, scope and objectives, and audience; it also lays out key differences between WHO guideline development, WHO prequalification, and other regulatory approval processes.

Section 2 explains the methodology used to develop the document.

Section 3 provides important background information – it outlines key steps in the WHO guideline development process and the framework that is used to assess evidence. This section also explains in general terms what evidence is sought by WHO to support guideline development, what may constitute "high certainty" evidence on benefits and harms of a new intervention, and how judgements are made during guideline development group meetings that are used to formulate WHO recommendations.

Section 4 is the core of the document. It can be read independently of the rest of the document if this is the area of primary interest. The section provides 28 key messages on how evidence should be generated to optimally inform WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections guideline development for TB diagnostics, including suggestions on key study protocol elements and guidance on what evidence may be generated beyond diagnostic accuracy and associated benefits and harms.

Section 5 describes processes led by departments within WHO other than the Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections that may be considered during the development of new TB diagnostics (e.g. WHO prequalification and WHO Essential Diagnostics List).

A note on language

The terminology used to describe evidence, outcomes and study design for diagnostics is overall less standardized and thus more variable than for therapeutics. We therefore provide a glossary to be clear what is meant within this document when certain terminology is used. We acknowledge that alternative terminology and definitions exist.

Diagnostic test, testing strategy, index test, diagnostic intervention…

Depending on the context, we may be using different terminology when referring to what is being evaluated. Within a study trying to determine the diagnostic accuracy of a new test, or a systematic review of such studies, the test under evaluation is typically referred to as the ***index test***. Within this document we also sometimes use this term, specifically in the context of diagnostic accuracy studies.

In general, during WHO policy development we are interested in understanding the effects of a certain intervention on health among other considerations. When thinking of diagnostics, the intervention here may represent use of a new test (the index test) or use of a new specimen type, use of a new diagnostic strategy (e.g. combining different tests in a certain way) or use of an existing test within a new delivery model. From that perspective, we are often using the term ***diagnostic intervention*** within this document, as a broader and more general term from the policy-making perspective.

Glossary

NOTE: Terms that appear in this glossary and are themselves defined in this section are highlighted as *italic and underlined*.

Testing terms

Reference standard: Is the test (or combination of tests) used to classify patients as having or not having the target condition. The reference standard is a measurement tool used to define sensitivity and specificity, not necessarily a test to compare the index test to (unless the reference standard is standard of care) and thus is often distinct from a relevant *comparator test*.

Index test: The test under evaluation, sometimes just referred to as "test" or "diagnostic" in this document. We mainly use this term when the test is evaluated for its diagnostic accuracy in detecting a target condition, i.e. in the context of a *diagnostic test accuracy* study where test results typically are not used to inform clinical decision-making.

Intervention (or diagnostic intervention): For the purposes of this document, this is defined as the test or testing strategy under evaluation. This could be a new technology (also referred to as *index test*), or a novel testing strategy (e.g. using the test in a new population, in combination with other tests, new TB screening algorithm, using novel specimen types etc.), a combination of tests or strategies (such as pooled specimen testing) or new way of delivering a test. We mainly use this term when speaking about evaluation in the context of developing policies or if evaluation is done in a *diagnostic randomized controlled trial* or other design aiming to directly estimate the effect of an *intervention* on *intermediate* or *final outcomes*.

Comparator (or comparator test): A comparator test or strategy is the test or strategy reflective of current standard of care for routine clinical use in a given setting and/or the recommended test for use (based on international or local policy). In some instances, it may be identical to the reference standard test but often it is not, e.g. liquid culture is used as a *reference standard* but is not standard of care for TB detection in high-burden countries. In the context of a *diagnostic test accuracy* study, the purpose of a *comparator* is akin to the purpose of a control group in a randomized controlled trial as it permits direct comparison of outcomes between the new intervention and standard of care. In the context of a *Diagnostic randomized controlled trial*, the *comparator* is used to guide management in the control group.

Patient spectrum: The range of patient characteristics (disease spectrum, age, comorbidities, severity of symptoms, risk factors, referral pathway, etc.) included in a study population. The disease spectrum includes severity and manifestations included in a study (e.g., early vs late TB, paucibacillary vs smear-positive TB). Sensitivity, specificity, are not fixed test characteristics, but depend on the patient and disease spectrum (spectrum effect). Including only "clear-cut" cases inflates sensitivity/specificity, leading to spectrum bias.

Specimen: A specimen is any sample or material collected from a person for diagnostic, screening or research purposes. It can include sputum specimens (e.g., for respiratory infections), non-sputum specimens such as blood, urine, stool, swabs, or tissue, and even aerosol-based samples from exhaled breath. Specimens may also take the form of diagnostic images like chest X-rays, ultrasounds, or CT scans. In short, a specimen is any biological sample or image used to detect disease, guide treatment, or support research.

Study and Trial terms

Diagnostic test accuracy study: A study that evaluates how well a test correctly identifies or excludes a target condition by comparing the results of the *index test* with those of a reference standard, usually expressed in terms of sensitivity and specificity.

Diagnostic randomized controlled trial: A trial in which eligible participants are randomly allocated to groups that receive either a new diagnostic interventions or standard of care as *comparator* to support clinical management. Typically, the focus of such trials is on estimating the effects of the novel intervention on intermediate or final outcomes, not on the measurement of test outcomes.

Non-randomized studies of interventions: An observational study that aims to evaluate the causal effect of an *intervention* but does not use randomization to assign participants to intervention and control groups. This includes cohort studies, case–control studies, cross-sectional studies, and *quasi-experimental studies*.

Quasi-experimental studies: Subset of non-randomized studies of interventions that typically provide lower risk of selection bias and confounding due to specific design features (e.g. difference in differences, interrupted time series analysis, regression discontinuity, instrumental variable etc.).

Diagnostic before-after study: An observational design in which clinical management decision are compared before and after the results of a *diagnostic test* become available to the decision-maker to estimate the effect tests results have on clinical decision-making.

Comparative diagnostic accuracy study (also referred to as head-to-head comparison): Study that directly evaluates two or more diagnostic tests by applying them to the same group of patients or same specimens under similar conditions and with their *accuracy* measured against the same *reference standard*. This approach allows for a fair and simultaneous comparison of the tests' performance (e.g., sensitivity, specificity, turnaround time).

Concordance studies: Studies that compare two or more diagnostic tests by applying them to the same group of patients under similar conditions without the use of a *reference standard*. This approach does not allow for computing of unbiased estimates for sensitivity and specificity, only the percent “agreement” (positive percent agreement positive percent agreement and positive percent agreement). “Agreement” does not mean “correct” and measures of overall agreement are not in themselves a sufficient characterization of the performance of a test.

Outcome terms

Test outcomes: Outcomes that measure properties of a test, such as *diagnostic test accuracy*, time to result and indeterminate rates.

Diagnostic test accuracy: The degree to which a diagnostic test correctly distinguishes between patients with and without the target condition (i.e. tuberculosis), typically expressed through sensitivity and specificity.

Comparative diagnostic test accuracy: Main outcome of interest that can be derived in a *comparative diagnostic test accuracy study*. Composed of the difference in sensitivity and specificity between index test and comparator (with 95%CI around that difference, accounting for the paired nature of the data), with the accuracy of both measured against a *reference standard*.

Non-positive non-negative results: Group of test and instrument-related outcomes occurrence of which leads to test results being neither positive, nor negative, including e.g. rates of instrument failures, test failures, invalid results or indeterminate results.[1]

Test robustness: A test's ability to remain unaffected by variations in environmental conditions (e.g. temperature, humidity, dust), employment by users with varying levels of training or experience, and differing levels of adherence to test procedures.

Positivity-rate (diagnostic yield or simply yield): The positivity-rate of a test is the proportion of people in whom it indicates presence of the target condition among all people to whom testing was offered.

Intermediate outcomes: Outcomes that occur in a causal pathway between test outcomes and *final outcomes*, relating to consequences of test outcomes, which are not *final outcomes* (e.g. reductions in pre-treatment loss to follow-up, change in clinical decision-making).

Final outcomes: Outcomes that are either ultimate measures of the health status at an individual level (e.g. quality of life or mortality) or critical metrics relating to long-term population-level health outcomes (e.g. transmission, case detection rates, TB incidence).

Patient-important outcomes: Outcomes that matter most to patients and other persons affected by a recommendation. Patient-important outcomes are typically *intermediate or final outcomes*. These could either be reported directly by patients (known as patient reported outcomes) or not and may represent a subset of *patient-centred outcomes* (i.e. health outcomes that are, or have been, identified, defined, prioritized and interpreted in partnership with patients reflecting their individual values, preferences and lived experiences).

Values: The relative importance people place on health outcomes. Values affect the weighting of desirable and undesirable effects, potentially modifying a recommendation based on the balance of effects derived from studies on diagnostic tests.

Value of knowing: An *intermediate outcome* reflecting any consequence for the wellbeing of a patient, or their family members or carers, that arises directly through the knowledge or information obtained as a result of testing (e.g. the value of having a diagnosis confirmed by a test, even if clinical management is not affecting by the result), rather than as a consequence of changed clinical management and related effects on health outcomes.

Evidence assessment terms

Patient-important-outcome-based approach (also referred to as "end-to-end" or "direct" approach): Approach to the evaluation of a diagnostic intervention, where *patient-important outcomes* (either *intermediate or final outcomes)* are compared directly between patients who received the *index test*/*diagnostic intervention* to those who received the *comparator test*/standard of care (each being managed supported by the respective test results). Evidence supporting this direct approach could come from *diagnostic randomized controlled trials*, *quasi-experimental studies* or *non-randomized studies of interventions*, where test results are used to inform clinical decision-making and patients

[1] Evans SR, Pennello G, Zhang S, Li Y, Wang Y, Cao Q, et al. Intention-to-diagnose and distinct research foci in diagnostic accuracy studies. The Lancet Infectious Diseases. 2025 Aug;25(8):e472–81.

are follow-up beyond the encounter where a specimen is obtained. This approach provides direct evidence in the sense that the effect of using a test on health outcomes is captured directly.

Diagnostic-accuracy-based approach (also sometimes referred to as the "linked evidence" or "indirect approach"): Approach to the evaluation of a diagnostic intervention, where typically the principal source of evidence stems from diagnostic test accuracy studies. Evidence on the diagnostic accuracy is then combined with other evidence and assumptions to assess the expected effects on final outcomes when using the *index test*/*diagnostic intervention* versus *comparator test*/standard of care. This approach provides indirect evidence in the sense that the effect of using a test on final outcomes is captured by linking multiple different pieces of evidence and judgements to arrive at a best estimate on the effect of using a test on health outcomes.

GRADE approach: The GRADE approach stands for Grading of Recommendations Assessment, Development and Evaluation. It is a systematic and transparent method used to assess the *certainty of evidence* and the strength of recommendations in healthcare and public health. It is used to develop clinical practice guidelines, and other recommendations.

Evidence: Evidence is the best available information or findings derived from systematic observation or research, that helps answer a research question to inform decisions in public health, clinical practice and policy development.

Certainty of evidence: In the GRADE framework, certainty of evidence refers to how confident one can be that the effect estimate observed is close to the true estimate. The factors that can reduce the certainty of evidence are: *risk of bias,* *imprecision*, *indirectness* and *inconsistency*; factors that can increase the certainty of evidence are: large magnitude of effect, effect of plausible residual confounding, dose-response gradient.

Bias: Systematic error arising from design, conduct, or analysis of a study that results in an incorrect estimate of an estimates for an outcome.

Imprecision: Refers to uncertainty about estimates due to random error e.g. as a result of small sample size and reflected in wide confidence intervals.

Indirectness: Refers to the degree to which the evidence directly applies to the specific clinical question or context under consideration. Indirectness can arise due to differences between the guideline question and the available evidence in population, *intervention*, *comparators* or outcomes.

Inconsistency: In the GRADE framework, inconsistency is present when results from different studies show substantial differences in effect estimates (i.e., they are not similar in direction or magnitude), and these differences cannot be explained by known factors such as differences in study populations, *interventions*, *reference standards*.

Introduction

1.1. Background

WHO develops global policies for the use of tuberculosis (TB) diagnostics through a rigorous process that involves the systematic use of high-quality evidence as its basis[2]. Whenever possible, global policy should be grounded in evidence from randomised controlled trials (RCTs) that directly measure patient-important outcomes. Early in the life-cycle of novel diagnostic tests, however, such evidence is rarely available, due to the need to determine the accuracy and benefits and harms of new technologies before using them for clinical decision making. Therefore, initial policy decisions most often rely on studies of diagnostic test accuracy together with other cost, feasibility, acceptability, and equity evidence to make indirect judgements about the direction and strength of the effects from use of a new diagnostic intervention on patient-important outcomes. However, the type and timing of evidence that can optimally support the policy process is not always clear to those funding, developing, and evaluating new diagnostic technologies, specimen types, or strategies. This lack of clarity has led to the generation of research evidence for WHO's guideline development process that often suffers from one or several important limitations:

(i) Evidence is of low certainty due to risk of study bias, indirectness, inconsistency, imprecision, dissemination bias or a combination of these factors;
(ii) Evidence is limited to diagnostic test accuracy data with no or only limited evidence on patient-important (intermediate or final) outcomes;
(iii) There is no evidence available on other criteria relevant for policy decision-making such as on feasibility, acceptability, equity and cost;
(iv) Methods and results are inadequately or inconsistently reported, hampering the ability to combine and analyse findings from one or more studies and leading to uncertainties about the variability and quality of the evidence base; or
(v) Evidence is lacking completely for certain populations that often would benefit most from the intervention (i.e., children or people living with HIV) or certain specimen types that are essential for diagnosis of disease with significant morbidity and mortality risks (i.e., non-sputum specimens or extrapulmonary TB).

Several well-established tools exist for the reporting and assessment of diagnostic research, such as the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool and the STARD (Standards for Reporting of Diagnostic Accuracy Studies) checklist. The QUADAS-2 tool provides a standardized framework used to assess the risk of bias and applicability of diagnostic accuracy studies. It helps researchers and reviewers evaluate study quality systematically. The STARD checklist provides guidelines to ensure clear, complete, and transparent reporting of diagnostic accuracy studies. Its use improves study reproducibility, interpretation, and comparability. Although additional guidance on study design for diagnostic accuracy studies exists [3-7], it is limited in scope and outdated. Importantly,

[2] Evidence generation for development of health products: a practical guide for WHO staff. Geneva: World Health Organization; 2023 2023.

[3]Schumacher SG, et al. Guidance for Studies Evaluating the Accuracy of Sputum-Based Tests to Diagnose Tuberculosis. J Infect Dis. 2019 Oct 8;220(220 Suppl 3): S99-S107. doi: 10.1093/infdis/jiz258. PMID: 31593597; PMCID: PMC6782025.

3. Drain, et al. (2019). Guidance for Studies Evaluating the Accuracy of Biomarker-Based Nonsputum Tests to Diagnose Tuberculosis. *The Journal of infectious diseases*, *220*(220 Suppl 3), S108–S115. https://doi.org/10.1093/infdis/jiz356;

4. Georghiou, S. al(2019). Guidance for Studies Evaluating the Accuracy of Rapid Tuberculosis Drug-Susceptibility Tests. *The Journal of infectious diseases*, *220*(220 Suppl 3), S126–S135. https://doi.org/10.1093/infdis/jiz106;;

5. Nathavitharana, R. R., et al (2019). Guidance for Studies Evaluating the Accuracy of Tuberculosis Triage Tests. *The Journal of infectious diseases*, *220*(220 Suppl 3), S116–S125. https://doi.org/10.1093/infdis/jiz243;

6. Hamada Y, et al. Framework for the evaluation of new tests for tuberculosis infection. Eur Respir J. 2021 Aug 19;58(2):2004078. doi: 10.1183/13993003.04078-2020. PMID: 33479110; PMCID: PMC8374690.

7. MacLean, et al. (2024). Tuberculosis treatment monitoring tests during routine practice: study design guidance. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, *30*(4), 481–488. https://doi.org/10.1016/j.cmi.2023.12.027

these resources do not provide specific guidance on diagnostic interventions for TB or provide clear guidance on needs for WHO policy recommendations. As new tests, sample types, strategies, and service delivery models for the detection of TB are developed, best practices for intervention evaluation need to be matched with tailored study considerations so that the evidence generated can maximally inform policy and benefit people around the world.

There is therefore a pressing need for clear, up-to-date direction from WHO to support donors, innovators, test developers and investigators in generating high-quality evidence that supports strong recommendations for impactful diagnostic interventions, especially as new specimen types, diagnostic strategies and technologies emerge.

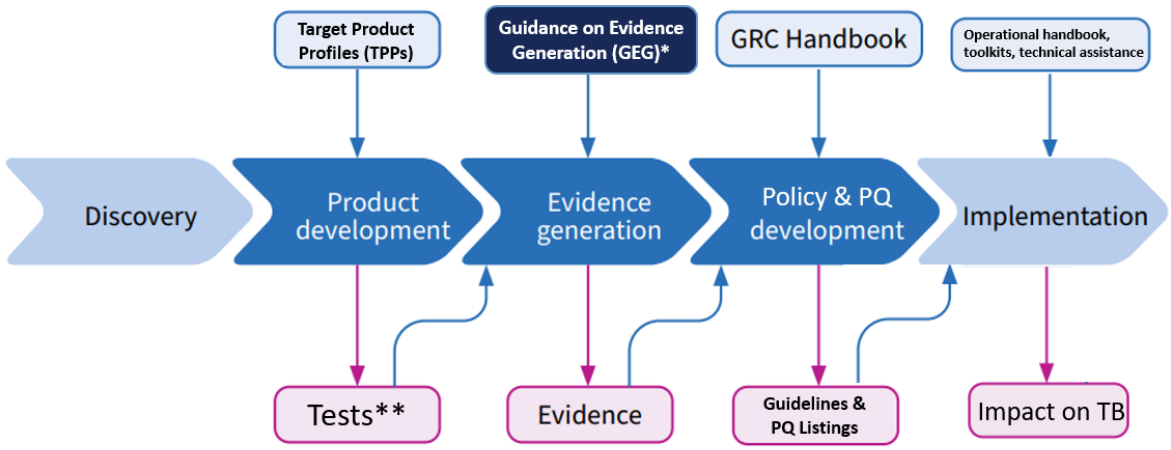
Beyond the specific study design considerations that are the focus of this document, active and early engagement of key stakeholders, particularly national TB programmes and affected communities, is critical to ensure that research outputs meet the needs of end users.

1.2. Purpose, scope and objectives

WHO guides the prioritization and development of new health products through the publication of target product profiles (TPPs), which outline desired product characteristics—such as intended use, target populations, safety, and efficacy. These TPP documents guide industry research and development, support regulatory submissions, and serve as planning tools for public health stakeholders. A list of current TPPs is provided in Annex 1 (Table A1). This document ("Guidance on Evidence Generation", GEG) is a critical complement to these TPPs. It aims to facilitate the production of high-certainty evidence to support the development of WHO policies on TB diagnostics, conducive to strong recommendations that are more likely to be implemented.[4] Figure 1.1 illustrates the role of the GEG as part of other WHO processes and products supporting steps in the TB diagnostic value chain.

[4] Nasser SMU, Cooke G, Kranzer K, Norris SL, Olliaro P, Ford N. Strength of recommendations in WHO guidelines using GRADE was associated with uptake in national policy. Journal of Clinical Epidemiology. 2015 Jun 1;68(6):703–7.

Figure 1.1 WHO processes and products supporting steps in the TB diagnostic value chain



* GEG informs evidence generation process for WHO guideline development; Technical Specification Series (TSS) documents and other PQ guidance inform evidence generation for PQ evaluations
** New "products" or diagnostic interventions may also include new specimen types, testing strategies or other diagnostic approaches

GEG: guidance on evidence generation; GRC: Guideline Review Committee; TPP: target product profile; WHO: World Health Organization. The chevron process shows a simplified "discovery-to-implementation value chain"; the red boxes below the chevron show the outputs of some of the steps in this process, which then feed into the next step; the blue boxes above the chevron show guidance documents that inform some of the steps in this process.

The core scope of this GEG document relates to generating initial evidence for WHO to assess first-in-class technologies, new diagnostic interventions (i.e., specimen types, strategies), or application of existing technologies to new populations (i.e., children, people with presumptive extrapulmonary TB, or people living with HIV) where the intended use is the detection of TB disease and resistance to anti-TB medicines. However, many of the principles outlined here will apply to tests for infection, tests for screening and tests for treatment monitoring. Key areas where needs may differ for these specific indications are provided in Annexes 8, 9 and 10, respectively. Further, while the scope of this document is tailored tuberculosis diagnostics, some of the principles may apply to evidence generation for diagnostic policy evaluations, in general. Detailed guidance on conducting diagnostic RCTs or other comparative-effectiveness designs to measure the effects of diagnostic interventions on intermediate and final outcomes is outside the scope of this document although high-level guidance is provided in section 4.5 (with some further references in Annex 5).

The objectives of this document are to:

1. Describe the key steps of the WHO guideline development process, including the application of the GRADE framework to diagnostics (Section 3);
2. Provide practical, study-level guidance on how to plan, conduct and report research on the potential health benefits and harms of new TB diagnostics, with specific attention to diagnostic test accuracy studies but also end-user values, resource requirements, cost-effectiveness, equity, acceptability and feasibility (Section 4); and
3. Serve as a consolidated, high-level reference on WHO guideline development, prequalification and other WHO processes relevant for the introduction of new TB diagnostics (Section 5).

1.3. Audience

This document is aimed primarily at stakeholders and organizations involved in generating evidence on new tuberculosis diagnostics, including commercial diagnostics developers, researchers, funders of such research and organizations involved in advocacy for funding and use of appropriate study

design. The document will also be informative for GDG members because it describes critical issues that are frequently discussed in GDG meetings as well as definitions, steps, and inter-step connections within the GRADE process that are often unknown to new members.

Specifically, this document may serve as a reference for:

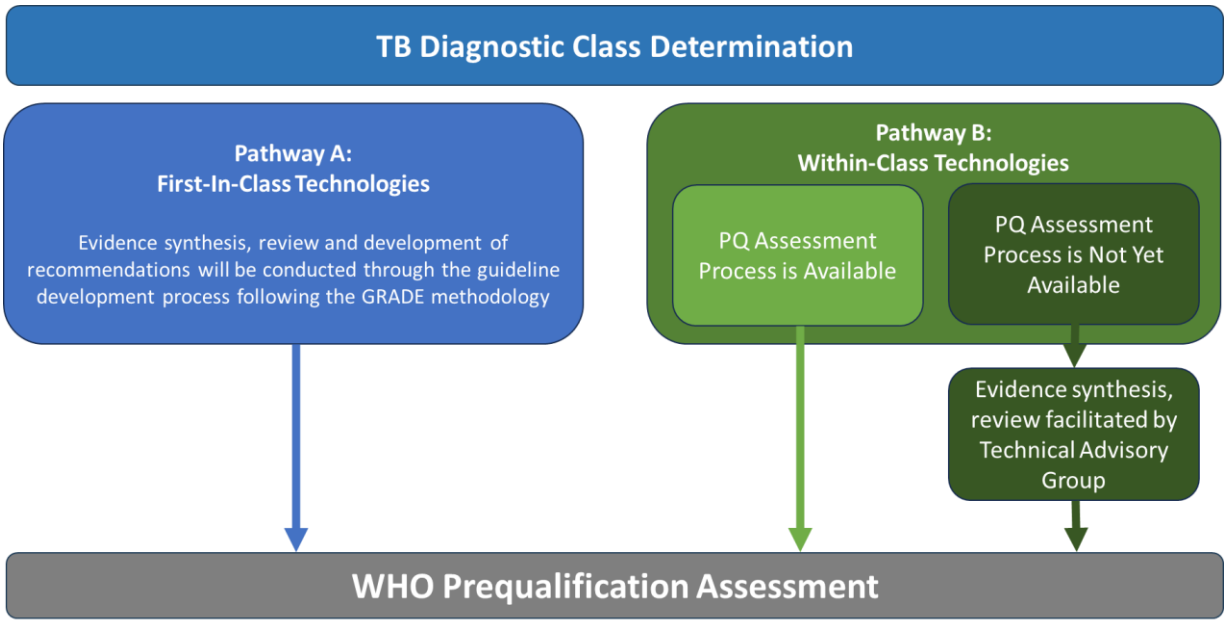
- **Ministries of Health** to understand how new and existing recommendations on TB diagnostic testing are generated by WHO, including which components they may lead evidence generation for, and which may be useful for national review or regulatory processes.
- **Test developers** to solidify understanding of how their product(s) will be assessed using which types of evidence from early evaluations;
- **Evidence generators, systematic reviewers and other researchers** to guide the design, execution, and reporting of independent research that can optimally inform the development of new TB diagnostic guidelines;
- **Donors** to understand how investments in evidence generation and assessment may be directed to maximally inform advances in global TB diagnostic policy; and
- **Guideline development group members** to gain a high-level understanding of the overall guideline development process and the critical role their contributions play at each step

1.4. WHO processes for guideline development and related processes

WHO's approach to developing policy on TB diagnostics has shifted from evaluating individual products to assessing classes of technologies. Once recommendations for a class are issued, products within that class undergo prequalification review for quality, safety, and performance through dossier submission, site inspection, and, where relevant, performance evaluation (see Sections 5.1). A new parallel assessment process is being piloted to reduce timelines for policy recommendations and procurement; if successful, it will become the standard pathway for TB diagnostics evaluation.

The process for policy development and PQ listing depends on whether a class for a new test already exists and whether a PQ process for the assessment of products in this class is already available or not. The who main pathways are outlined in Figure 1.2. Further information on the function of WHO Technical Advisory Group within this pathway is provided in section 5.2.

Figure 1.2 WHO processes for assessment of new TB diagnostics



The WHO assessment process for TB diagnostics has recently evolved to focus on evaluating classes of TB diagnostic technologies rather than specific products. Class determination is managed by WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections for new diagnostic testing technologies, and it includes an evaluation of various characteristics such as complexity of testing procedure, principle, infrastructure and human resources needed, etc. (Ref: WHO consolidated guidelines on tuberculosis. Module 3: diagnosis. Geneva: World Health Organization; 2025. Licence: CC BY-NC-SA 3.0 IGO)**. Pathway A**: "First-in-class" diagnostics – For technologies that differ significantly from existing classes (e.g., new principles of action, specimen types, or testing strategies). These undergo evidence synthesis, review using the GRADE methodology, and guideline development. If recommended, technologies are added to WHO consolidated guidelines and referred for prequalification, once eligible. Until then, WHO/GTB recommendations remain valid. **Pathway B:** Within-class diagnostics – For technologies matching an existing class. If eligible for prequalification, manufacturers may apply directly. If not, evidence is reviewed by the WHO disease programme and the Technical Advisory Group on TB Diagnostics and Laboratory Strengthening. If class recommendations are deemed applicable, WHO issues a policy statement and adds the technology to the relevant class in its policy guidance. This recommendation also stands until prequalification is completed. If there is a negative recommendation from the prequalification assessment, the WHO recommendation will change accordingly to reflect this.

A separate interim, time-limited mechanism is the Expert Review Panel for Diagnostics (ERPD). It aims to facilitate early access to innovative diagnostics that may have a substantial public health impact, but are not yet recommended by WHO, are not in the scope of prequalification or have not yet been prequalified or undergone stringent regulatory assessment by a founding member of the Global Harmonization Task Force (see Section 5.3). The WHO Essential Diagnostics List (EDL)[5] is an evidence-based register of IVDs that supports countries to facilitate their decision-making processes for selection and procurement of diagnostics (see Section 5.4). The WHO Coordinated Scientific

[5] World Health Organization. The selection and use of essential in vitro diagnostics: report of the fourth meeting of the WHO Strategic Advisory Group of Experts on In Vitro Diagnostics, 2022 (including the fourth WHO model list of essential in vitro diagnostics). Geneva: World Health Organization; 2023. (WHO Technical Report Series; no. 1053). 9789240081093-eng.pdf

Advice (CSA) procedure is a single-entry service that lets developers of diagnostics (and other priority health products) obtain a joint, written assessment of their development plans from both the relevant WHO technical department and the WHO PQ Team (see Section 5.1).

There are important similarities and differences between the evidence needs for WHO recommendations, WHO prequalification and regulatory approval of new TB diagnostics. This document pertains to the evidence needs for WHO policy development but also makes some reference to prequalification and regulatory requirements, where relevant. Table 1.1 provides an overview of the key differences in scope and approach between these processes. Further details on the remit, approaches, quality assurance measures, and post-achievement support for each process are available in Annex 1. Of note, while the approach to sourcing of evidence may differ between these processes, the same underlying analytical and clinical studies may generate evidence for multiple processes at the same time.

Table 1.1. Overview of WHO disease programme assessment, WHO prequalification, and regulatory approval for TB diagnostics

	WHO disease programme assessment	WHO prequalification assessment	National regulatory approval
Triggered by	Identified global public health need with developed, design-locked and market accessible products	Diagnostic class covered by a WHO recommendation and identified as eligible for PQ by the disease programme through prior disease programme endorsement or determination of new product listing within an existing class.	National regulatory authorities or designated bodies have the mandate to assess medical devices, including diagnostics, and authorize their placing on the market.
Scope	Classes of TB diagnostic technologies	Specified product brands	Specified product brands
Source of evidence	Systematic review reports and summaries of evidence from published and final, locked, and quality unpublished trials and studies	Product and manufacturing/QMS related evidence typically submitted in the marketing authorization application. Independently generated reports (such as inspection reports, evaluation reports etc.) reflecting conformity assessment by the regulatory authority /WHO or designated body.	
Focus of evidence assessment	Assessment of diagnostic class impact on patient important outcomes, diagnostic accuracy, economic evidence, feasibility, accessibility and equity aspects of technologies within a diagnostic class in specific patient populations against an appropriate comparator.	Assessment of product safety, performance and quality[6], including labelling, quality management and manufacturing	

[6] International Medical Device Regulators Forum. (2024). *IMDRF Code IMDRF/GRRP WG/N47 FINAL:2024 (Edition 2)*. IMDRF. https://www.imdrf.org/

Outcome	WHO recommendations for diagnostic classes or referral to prequalification for products in existing classes	Prequalification listing of product brands	Regulatory listing of product brands
Meaning of a decision	A recommendation for the class of technology would make the included tests in the evaluation eligible for Global Fund grants and procurement via GDF, UN agencies, governments and other donors.	UN agencies, international or intergovernmental procurement organizations and/or WHO Member States may use WHO's list of prequalified IVDs to inform their respective procurement decisions.	A test is deemed licensed or approved for the purposes of its importation, sale or advertisement within a certain jurisdiction.

2. Methodology for development of GEG

The process of developing this guidance document on evidence generation involved several structured steps, based on a standardized process (1), as outlined below.

Step 1: Establishment of the Steering Group

A steering group, consisting of WHO staff from relevant WHO headquarters departments and regional offices, as well as WHO Prequalification was formed. The core role of the steering group was to oversee the scope of the planned WHO GEG and support the administrative process of its planning, development, review, publication and dissemination.

Step 2: Establishment of the Scientific GEG Development Group

The Scientific GEG Development Group (SGG) was formed; it comprised leading trialists, scientists, public health officials, regulators, economists, social scientists, end users, civil society representatives, individuals with lived experience and experts involved in developing WHO policy recommendations for TB diagnostics. The SGG played a pivotal role in supporting the entire development process of the GEG. The contributions of SGG members included reviewing drafts at various stages, participating in discussions during meetings, and providing direct input into the drafting process. The standard WHO procedures for declaring conflicts of interest were adhered to for all members of the SGG (listed in the Acknowledgements).

Step 3: GEG document and review

The initial draft of the GEG document, referred to as version 0, was developed by the WHO Secretariat. This version served as the foundation for subsequent revisions and stakeholder consultations. The SGG reviewed version 0 and provided detailed written feedback. Based on the consolidated inputs received, the document was revised, resulting in an updated version (version 0.1). One additional iteration of input by the SGG and revision resulted in version 0.2.

Step 4: Public comment and external review

To ensure input from stakeholders not represented in the SGG (owing to possible conflicts of interest), and to facilitate the broadest possible input, version 0.2 was made publicly available and public comment invited. The feedback received was considered during discussions with the SGG and revision of the document. In parallel, a group of external reviewers with experience in TB diagnostic studies and guideline development was asked to provide an independent written review of draft version 0.2. Lastly, version 0.2 was also shared again with the SGG for their review in preparation of the consensus meeting.

Step 5: Consensus meeting and finalization of the document

A consensus meeting was convened to resolve any remaining questions on version 0.2, including those raised by external reviewers, funders, industry and other stakeholders. Following this meeting, version 1 of the GEG document was finalized and prepared for dissemination.

3. WHO guideline development process

3.1. WHO guideline development for TB diagnostics using the GRADE approach

The fundamental means through which WHO fulfils its technical leadership in health are review of evidence and development of normative products such as guidelines; (2). The process for developing WHO guidelines is detailed in the *WHO handbook for guideline development (*7). A short, open-access online course on the use of GRADE to develop WHO guidelines is available (8).

What is a WHO guideline?

A WHO guideline is any document developed by WHO that contains recommendations for clinical practice or public health policy. A recommendation tells the intended end users of the guideline what they can or should do in specific situations, individually or collectively, to achieve the best health outcomes possible. It offers a choice among different interventions or measures expected to have a positive impact on health and implications for the use of resources.

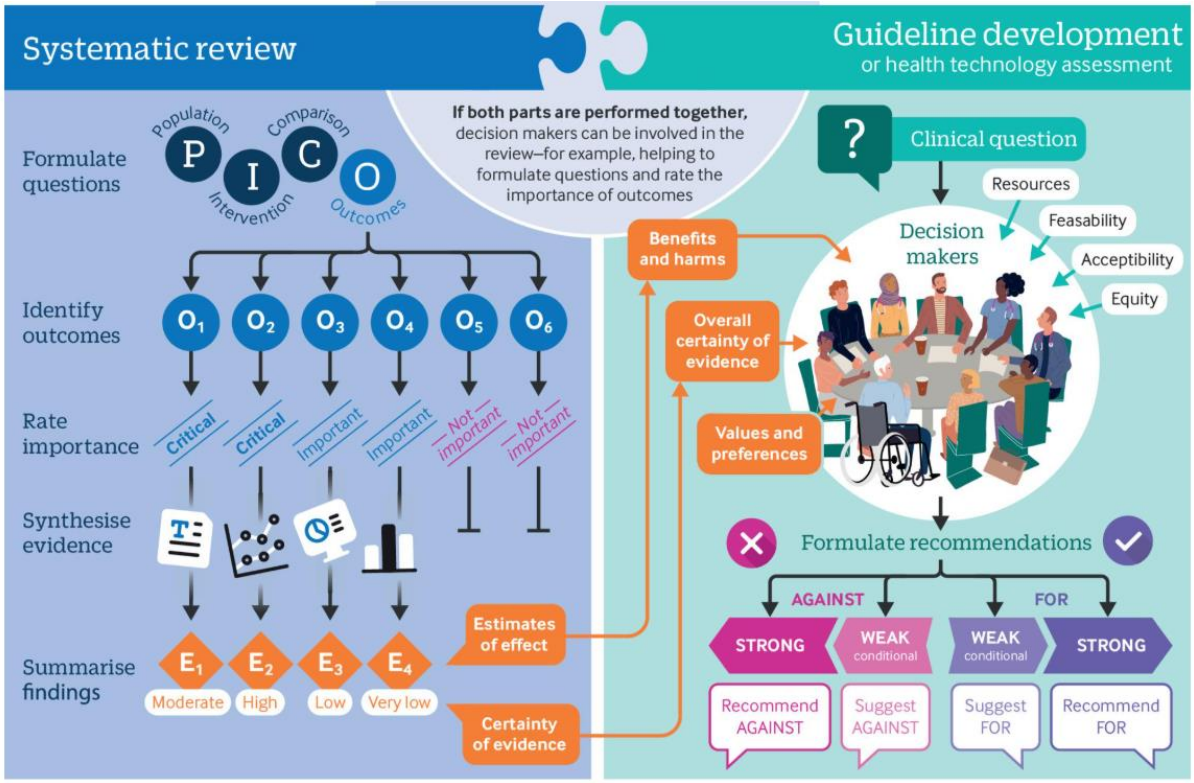
WHO uses the internationally recognized GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach to assess the certainty of a body of evidence, and to develop and report recommendations (3-7). Key principles for the development of WHO guidelines include:

- ✓ explicit, inclusive and transparent processes for developing recommendations (i.e. users can see how and why a recommendation was developed, by whom and on what basis);
- ✓ use of standardized, transparent processes and methods in each step of guideline development to minimize the risk of bias in and increase the applicability of the recommendations; and
- ✓ recommendations developed based on a systematic and comprehensive assessment of the balance of an intervention's potential health benefits and harms, and explicit consideration of other relevant factors (Table 3.2).

This section provides a brief overview of critical steps in the process that WHO uses to assess evidence for policy development (Figure 3.1), including:

- development of the scope and recommendation questions using the PICO (population, intervention, comparator and outcome) format to guide evidence retrieval and synthesis (Section 3.2);
- evaluation of the "certainty" of the evidence and preparation of evidence profiles (Section 3.3);
- the two approaches to the assessment of diagnostic interventions (Section 3.4);
- how decisions are made across evidence-to-decision (EtD) criteria (Section 3.5);
- formulation of the recommendations (Section 3.5).

Figure 3.1 WHO process for systematic review and guideline development using the GRADE approach



Source: [BMJ 2025;389:e081903](BMJ 2025;389:e081903) (adapted; pending permission for use)

3.2. Developing the scope and recommendation questions using the PICO format to guide evidence retrieval and synthesis

One of the critical initial steps in the development of a guideline is the definition of the scope and formulation of recommendation questions in the PICO (**P**opulation, **I**ntervention, **C**omparator, **O**utcomes) format. These are developed in response to a public health need for which a new diagnostic test has been developed, and evidence has been generated. The formulation process considers the role of the new technology, specimen type, or testing strategy (i.e., initial detection of TB, follow-on detection of drug resistance), how it fits into an overall clinical pathway through use of an analytical framework (Section 3.3.2), and which outcomes it will impact[7]. Questions are drafted by the WHO Secretariat and then reviewed, revised, and finalized with inputs from the WHO steering committee and guideline development group and confidential outcome ranking by the guideline development group.

Did You Know? Recommendation questions in the PICO format are core to the guideline development process as they are used to determine which evidence should be collected, how it is synthesized, how the findings of evidence assessment are reflected for decision makers, and how decisions based on the evidence are reflected in new recommendations.

[7] Neumann I, Souza-Pinto B, Meerpohl J, Dahm P, Brennan S, Alonso P, et al. Making answerable questions. In: Neumann I, Schünemann H, editors. The GRADE Book version 10 (updated September 2024): The GRADE Working Group; 2024.

Once defined, systematic reviews and evidence syntheses for each recommendation question are typically commissioned through independent researchers (i.e., those not involved in the generation of evidence). If only a single study or trial provides pertinent evidence for the recommendation question, the evidence review will focus on that study or trial. Detailed guidance on the performance of systematic reviews is provided in the *WHO handbook for guideline development* [8] and elsewhere – for example, in the *Cochrane Handbooks* [9] – and is beyond the scope of this document.

3.3. Evaluating the certainty of evidence and preparing evidence profiles

Once the evidence has been retrieved and synthesized through a systematic review, a critical next step is the assessment of the certainty of evidence (in the past this was also referred to as quality of evidence). In the context of evidence syntheses, the certainty of the evidence is defined as the "certainty that an estimate of association or effect is correct or, better, that a true effect lies on one side of a specified threshold or within a chosen range [10,11,12]. In the context of guideline development, the certainty of the evidence reflects the confidence that the estimates of an effect are adequate to support a particular decision or recommendation. GRADE includes four levels of evidence certainty (high, moderate, low or very low). There are five domains that could affect the certainty of evidence of test accuracy: risk of bias, indirectness, inconsistency, imprecision and dissemination bias (see Annex 3 for detailed descriptions). If important uncertainty is identified based on these domains, it is reflected by a downgrading of the evidence domain by one or more categories (i.e., from high to moderate, low or very low). There are three factors which could also lead to upgrading of the certainty of evidence: large magnitude of effect, effect of plausible residual confounding, dose-response gradient, however, we are not aware of any precedence for upgrading in the context of diagnostics. For qualitative evidence, GRADE CERQual is a transparent and structured approach for assessing how much confidence to place in individual review findings (i.e., to assess the extent to which the review finding is a reasonable representation of the intervention). Details about this tool are provided in Annex 3.

Once certainty of the synthesized evidence is complete, the systematic review teams format the findings of their analyses into so called 'evidence profiles', which display the summary results from a systematic review together with the certainty of evidence ratings.

3.4. The two approaches to the assessment of diagnostic interventions

Broadly two approaches exist for generating evidence on the effects of new diagnostics on health and thus the type of evidence available to support guideline development: the diagnostic-accuracy-based approach, and the patient-important-outcome-based approach (see definitions Glossary, box below and Figure 4.1).

[8] WHO handbook for guideline development, 2nd ed. 2nd ed. Geneva: World Health Organization; 2014.

[9] Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. 1st edition. Chichester (UK): John Wiley & Sons, 2023

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

[11] Schunemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol. 2016;75:6–15.

[12] Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401–6.

The two approaches to the assessment of diagnostic interventions
[Note: see more technical definitions in the Glossary on page x]

The patient-important-outcome-based approach looks directly at how a test affects people's health. It compares health outcomes between patients who are managed using one test and those managed using another test or usual care. The evidence come from studies where doctors use the test results to guide treatment decisions and then follow patients over time to see what happens.

Diagnostic-accuracy-based approach first looks at how well a test correctly identifies people with or without a disease, and then uses that information to estimate how the test might affect health outcomes. It combines results from test accuracy studies with other types of evidence and assumptions — how test results influence treatment decisions and how effective those treatments are.

Given the core scope of this guidance is to provide key messages on how to generate initial evidence on design-locked assays, the primary focus of this document will be on generating evidence to support the diagnostic accuracy-based approach for WHO guideline development. However, it is helpful to understand both approaches, their respective advantages and roles. In general, the patient-important-outcomes-based approach is preferred over diagnostic-accuracy-based approach because of the uncertainties when trying to predict whether or to what degree the introduction of a new test with certain diagnostic accuracy will ultimately affect final health outcomes. However, evidence for use in the patient-important-outcome-based approach can typically only be generated once a test has been recommended for use. Furthermore, generating evidence on a tests' effects on patient important outcomes may be more complex, more costly, may be less generalizable, and have other limitations (see table 3.3). Therefore, initial WHO policy recommendations for a new diagnostic intervention are typically made using the diagnostic-accuracy-based approach.

Table 3.1. Advantages and use of diagnostic-accuracy-based and patient-important-outcome-based approach to the assessment of evidence on desirable and undesirable effects (see also Section 4.5.1)
[NOTE: This is an early draft with internal discussions not concluded but input welcome]

	Diagnostic accuracy-based approach	Patient-important-outcome-based approach
STUDY DESIGNS	Primarily diagnostic accuracy studies, ideally supported by studies facilitating a linked-evidence approach to enable extrapolating from diagnostic accuracy to presumed effects on patient-important outcomes.	Ideally diagnostic randomized control trials on the effects of diagnostic interventions on intermediate and final outcomes. Quasi-experimental designs, and other non-randomised studies of interventions may also provide supportive evidence.
ADVANTAGES	• Can be generated at early stages in the life-cycle of a test, e.g. before a new test is in clinical use • Less resource intensive and quicker, thus typically broader evidence base available • May pose fewer challenges with generalizability than patient-important-outcome-based	• Evidence is more direct, requiring fewer assumptions on the intervention's effect on patient-important outcomes • Does not require testing with a reliable reference standard test • May be better suited to address broader questions about diagnostic interventions (e.g. effects of changing

	approach because it specifically focuses on test outcomes (and the additional variability in how test results affect downstream outcomes is not captured)	how diagnostics are delivered or of testing a wider population)
WHEN TO USE	The diagnostic accuracy-based approach can and should be used early in the life-cycle of a test. Its use is most appropriate if • The test is not yet recommended, and performance is just being established (and thus the test cannot be used to guide clinical decisions) • A good and feasible reference standard exists • The detectable patient spectrum is largely unchanged vis a vis the testing strategies that were used in trials to evaluate treatment effects	The patient-important-outcome-based approach can typically only be used after test performance characteristics have been established so that it can be employed to guide clinical decision making. Generating evidence to support this approach is most important when the diagnostic intervention may lead to significant changes in program implementation or patient spectrum, such as: • how and where testing is performed (e.g., from centralized labs to point-of-care) • changes in the eligible population being tested (e.g. testing asymptomatic or lower-risk individuals that would require a broader approach to patient enrolment), resulting in a possible change in the patient spectrum being diagnosed This approach can also be used when a suitable reference standard is not available, since use of a reference standard test is not required in studies contributing evidence (see also Section 4.5.1)

Further guidance on what to consider when deciding which approach to take is provided in sections 4.1 and 4.2. Additional information on implications for how evidence is reviewed depending on the approach taken is described in section 3.5.

3.5. Evidence to Decision framework

The WHO uses an Evidence to Decision (EtD) framework to guide the development of health recommendations by providing a structured format for deliberation and for moving from evidence to recommendations. After evidence is gathered and assessed for certainty, the GDG panel reviews the findings in this structured format to ensure transparent, evidence-informed decision-making. When the diagnostic-accuracy-based approach is used, typically seventeen criteria are used to guide deliberations on accuracy of the diagnostic intervention, balance of effects (i.e. balance of benefits and harms), cost and cost-effectiveness, acceptability, feasibility and equity. The decisions are made by consensus (preferred) or voting (if needed), recorded and made public together with the guideline text. Table 3.4 explains the 17 EtD criteria typically evaluated as part of the overall assessment of evidence. If evidence is available to use the patient-important-outcomes-based approach a slightly shorter and simpler EtD framework with 12 EtD criteria is used (the five criteria omitted are highlighted with an asterisk in Table 3.2).

Table 3.2. Overview of the 17 EtD criteria typically evaluated as part of the overall assessment of the evidence

EtD criterion (GEG section to reference for further detail)		**Signalling questions**	**Explanation and elaboration**
1	**Problem**	Is the problem a priority?	Providing background on whether and why the problem (i.e., tuberculosis) is a priority. In TB, GDGs have consistently judged this as 'yes.'.
2 (4.3)	**Test accuracy***	How accurate is the test?	Summary estimates of diagnostic test accuracy are derived from systematic reviews and meta-analyses using one or more reference standards, and, where applicable, compared to the accuracy of comparator tests.
3 (4.3, 4.4, 4.6)	**Desirable effects**	How substantial are the desirable effects?	Judgement on how large the **desirable** effects of the intervention are, based on outcomes where the diagnostic intervention shows **better** results than the comparator. Typically, the number of **true positives** and **true negatives** are displayed alongside other test and intermediate outcomes (e.g. more rapid time to diagnosis, lower rate of indeterminate test results, increased diagnostic yield etc.).
4 (4.3, 4.4, 4.6)	**Undesirable effects**	How substantial are the undesirable effects?	Judgement on how large the **undesirable** effects of the intervention are, based on outcomes where the diagnostic intervention shows **worse** results than the comparator. Typically, the number of **false positives**** and **false negatives**** are displayed alongside other test and intermediate outcomes (e.g. slower time to diagnosis, higher rate of indeterminate test results etc.).
5 (3.3)	**Certainty of the evidence of test accuracy***	What is the overall certainty of the evidence of test accuracy?	Judgment based on review of the risk of bias, indirectness, inconsistency, imprecision and dissemination bias associated with the reviewed evidence. These domains are closely aligned with assessments made and proposed by the systematic review team to the GDG panel for deliberation. A separate assessment is conducted for both test sensitivity and specificity, and the overall certainty of evidence is determined by taking the lower of the two certainty ratings (i.e., the lowest certainty rating between sensitivity and specificity guides the final judgment).
6 (3.3, 4.4.3)	**Certainty of the evidence of test's effects***	What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?	Judgement about how confident the GDG panel is that the diagnostic intervention leads to direct benefits, adverse effects or burden of the intervention, i.e. not as a result the subsequent management but through the testing process itself. In the context of TB tests, one typically needs to consider only possible procedural harms or inconveniences associated with specimen provision, (e.g. as a result of obtaining a specimen).

7 (3.3, 4.5)	**Certainty of the evidence of management's effects***	What is the overall certainty of the evidence of effects of the management that is guided by the test results?	Judgement about how confident the GDG panel is that the treatment or management improves health outcomes once the test guides that particular management/treatment. In the case of TB diagnostic intervention, it usually means how confident is the panel that the appropriate TB treatment (i.e., TB preventive, disease, or drug resistance regimens) will be beneficial based on the results of the intervention. Relevant WHO recommendation and their strength is considered here.
8 (3.3, 4.5)	**Certainty of the evidence of test result/management***	How certain is the link between test results and management decisions?	Judgement on how confident the GDG panel is that test results will inform and affect management of patients, including whether the results at the given diagnostic accuracy values would be used to guide treatment decisions. Rapid turnaround time and good interpretability of results can improve linkage and help reduce barriers to patients receiving the appropriate treatment after obtaining a test result.
9 (3.3)	**Certainty of effects**	What is the overall certainty of the evidence of effects of the test?	All of the above judgements on the certainty of evidence criteria (criteria 6-8) are reviewed to determine the overall certainty of the effects of the intervention, management from the intervention and test result management. The overall certainty rating across outcomes for a recommendation is typically based on the lowest certainty of any outcome deemed critical for the decision.
10 (4.7.1 and A4.4)	**Values**	Is there important uncertainty about or variability in how much people value the main outcomes?	This question refers to evidence on how much people value the outcomes for which evidence is available and if these values would differ based on population, age group, gender, sex, and other relevant subgroups.
11 (A3.6)	**Balance of effects**	Does the balance between desirable and undesirable effects favour the intervention or the comparison?	The balance of effects reflects the risk–benefit ratio of an intervention, considering the overall certainty of the evidence and how the outcomes are valued by those receiving it. It is thus based on a review of the judgements on the previous four EtD criteria (i.e., criteria 3, 4, and 9; desirable effects, undesirable effects, certainty of effects and values).
12 (4.7.2)	**Resources required**	How large are the resource requirements (costs)?	Assessment of the overall costs (direct and indirect) for implementing the diagnostic intervention compared to the current standard of care (i.e., the comparator).
13 (3.3)	**Certainty of evidence of required resources**	What is the certainty of the evidence of resource requirements (costs)?	Judgement about certainty of the evidence synthesized on resources required for the intervention.
14 (4.7.3)	**Cost–effectiveness**	Does the cost–effectiveness of the intervention favour the intervention or the comparison?	Assessment of whether the diagnostic intervention is cost-effective compared to the current standard of care (i.e., review of evidence from systematic reviews and modelling exercises in more than one intended setting of use).

15 (4.7.4)	**Equity**	What would be the impact on health equity?	Assessment of whether the diagnostic intervention will reduce or worsen health inequities. Different socio-economic groups, age groups, sexes, genders, and geographies are considered against the practical characteristics and categories of evidence.
16 (4.7.5)	**Acceptability**	Is the intervention acceptable to key stakeholders, in relation to the comparator?	Assessment on whether the intervention is considered to be acceptable by key stakeholders. The judgement often relies on synthesized evidence specific to acceptability, which is affected by a multitude of factors, such as expected health benefits, timeliness of the result and ease of use.
17 (4.7.6)	**Feasibility**	Is the intervention feasible to implement, in relation to the comparator?	Assessment of whether implementing the diagnostic intervention in the intended settings of use is considered feasible by key stakeholders. This considers aspects like infrastructure needed to implement the intervention, costs, ease of use, training and human resource requirements, etc.

EtD: evidence to decision; GEG: guidance on evidence generation. Note: Annex 3.2 provides the options for judgements to be made across each of the 17 criteria.
* Additional EtD criteria used in the diagnostic-accuracy-based approach. In the patient-important-outcomes-based approach, these are omitted with the criteria otherwise identical.
**False positives can lead to unnecessary stress, fear and anxiety in people where a person does not actually have the condition, but the diagnostic intervention says the person has the condition. This may lead to unnecessary treatment, follow on tests, money and medical resources. False negatives imply that a test says a person doesn't have the condition when they actually do. This might lead to delayed treatments, transmission of the disease and false sense of security with wrong diagnosis.

3.6. Developing recommendations

Recommendations are developed based on the judgements made across the EtD criteria (Annex 3). Typically, four factors with the strongest influence on the direction (i.e., for or against) and strength (i.e., strong or conditional) of a recommendation are:

- the certainty of the evidence (Section 3.3);
- values and preferences related to the health outcomes (Section 4.7.1, Annex 4.4);
- the balance of benefits and harms (Annex 3.6); and
- resource implications (Section 4.7.2).

When taken together, the direction and strength of all EtD criteria define which of five types of recommendations may be made (see Figure 3.1):

- ***strong*** recommendation ***for*** the diagnostic intervention;
- ***conditional*** recommendation ***for*** the diagnostic intervention;
- ***conditional*** recommendation ***for*** either the diagnostic intervention or the comparison;
- ***conditional*** recommendation ***against*** the diagnostic intervention; and
- ***strong*** recommendation ***against*** the diagnostic intervention.

Table 3.2 provides explanation on the conditions that typically need to be met to make strong recommendations.

Table 3.2 Factors impacting the strength and direction of a recommendation

	A strong recommendation may be justified if:	**A conditional recommendation may be expected when:**
Overall confidence in effect estimates	There is high or moderate confidence in effect estimates (or in special circumstance when the confidence is low or very low)	There is low or very-low confidence in effect estimates
	AND	**OR**
Balance between benefits and harms	The benefits clearly outweigh the harms or vice versa	The balance between benefits and harms is close
	AND	**OR**
Uncertainty and variability in stakeholder values and preferences	All or almost all fully informed stakeholders (including patients) will make the same choice	There is variability or uncertainty in what fully informed stakeholders (including patients) may choose
		OR
	AND	
Resource considerations	The benefit of the intervention is clearly justified (or not) in all or almost all the circumstances	The benefit of the intervention may not be justified in some circumstances

Detailed explanations and implications of these recommendations is provided in Annex 3. For example, strong recommendations are most often supported by high certainty evidence with clear values of the health outcomes where the benefits greatly outweigh the harms of the new TB diagnostic

intervention, and its use is found to be affordable for disease control programs. Finally, the recommendation is presented in its final format, noting the strength of recommendation and certainty of the evidence (i.e., "*For adults and adolescents with signs or symptoms of TB or who screened positive for pulmonary TB, low-complexity automated NAATs should be used on respiratory specimen as initial diagnostic tests for TB, rather than smear microscopy or culture. (Strong recommendation, high certainty of evidence*)."

4. Guidance on evidence generation

Evidence refers to findings from research and other credible sources of knowledge used to inform decisions in public health, clinical practice, and policy development. To ensure that clinical recommendations and public health policies are meaningful, relevant, and actionable, it is essential to generate high-quality, policy-relevant evidence.

The generation of high-quality, policy-relevant evidence begins with a clear understanding of the clinical pathway, the role of the diagnostic test within it and how use of the test may improve patient-important outcomes (i.e. its value proposition). This involves identifying where the test fits within the continuum of care, how it influences clinical decision-making, and what specific benefits it is expected to deliver for patients, health systems, and populations (Section 4.1 below).

Evidence on the benefits and harms of a new test, specimen type or testing strategy (all referred to as *diagnostic intervention* for the remainder of the section for brevity), assessed using appropriate reference standards and compared to the current standard of care (i.e. comparator(s)), are a core component of WHO guideline development and are referred to as 'desirable' and 'undesirable' effects. Several approaches can be used to generate evidence on desirable and undesirable effects, including randomized controlled trials providing direct evidence on health or population-level outcomes, and combinations of diagnostic accuracy studies and other research findings providing indirect evidence on the impact of the intervention (see section 3.6).

The following sections provide further guidance on considerations related to

- defining the **value proposition** and use of an **analytical framework** to inform design and outcome selection (Section 4.1 and 4.2);
- guidance on evidence generation for **diagnostic test accuracy** (Section 4.3);
- guidance on generating **additional evidence as part of diagnostic accuracy studies** and to complement the diagnostic-accuracy-based approach (Section 4.4);
- guidance on generating **evidence to support linkage** across the analytical framework (Section 4.5);
- guidance on generating evidence on **patient-important outcomes** (Section 4.6).

All sections are guided by the analytical framework shown in Figure 4.1, which highlights relevant outcomes, and the evidence needed for their measurement. We note that researchers are not expected to deliver evidence relating to each of the 28 key messages in a single study. For example, a diagnostic test accuracy study may or may not generate additional evidence on values, preferences or costs.

4.1. An analytical framework to guide evidence generation on accuracy and health outcomes

An analytical framework is useful to identify the types of evidence needed to evaluate the effect of introducing or changing a test or testing strategy on health outcomes. A widely used example of such a framework was developed by USPSTF (U.S. Preventive Services Task Force) for the evaluation of

screening programmes,[13] which we adapted for the context of TB diagnostics (see Fig. 4.1). Its primary purpose is to ensure that the evidence generated and collected comprehensively captures the relationship between the intervention and its effects on health. It thus relates to the EtD criteria 2-11 within the diagnostic-accuracy-based approach and to the EtD criteria 2-6 (see Table 3.2 in Section 3.5); other EtD criteria (such as acceptability, cost etc.) may also be captured and considered in a broader logic model. The framework also displays the two principal approaches that can be taken to generate such evidence (i.e. the accuracy-based approach and the patient-important-outcome-based approach; see Glossary and Section 3.4). Conceptually, the framework maps the pathway from the target population to the ultimate health and population-level outcomes, outlining key steps in the causal chain. The process of linking these evidence components may be informal, such as through expert judgment by GDG panel, or formal, through decision-analytic models or other types of modelling approaches.

The framework shown in Figure 4.1 should be adapted depending on the specific target population and diagnostic intervention of interest. Numbers in the figure refer to key actions, processes and direct associations, leading to outcomes. The numbers (1-4) within the framework relate to key questions that are assessed when following the **diagnostic-accuracy-based approach:**

(1) How accessible is testing for patients (e.g. based on how decentralizable the test is or how easily specimens can be obtained)?

(2) How well does the test perform in terms of test outcomes?

(3) Do changes in test outcomes lead to changes in intermediate outcomes?

(4) Do changes in intermediate outcomes lead to changes in final outcomes?

The letters (A and B) within the framework relate to key questions that are assessed when following the **patient-important-outcomes-based approach:**

(A) Does the diagnostic intervention improve intermediate or final outcomes among the eligible population?

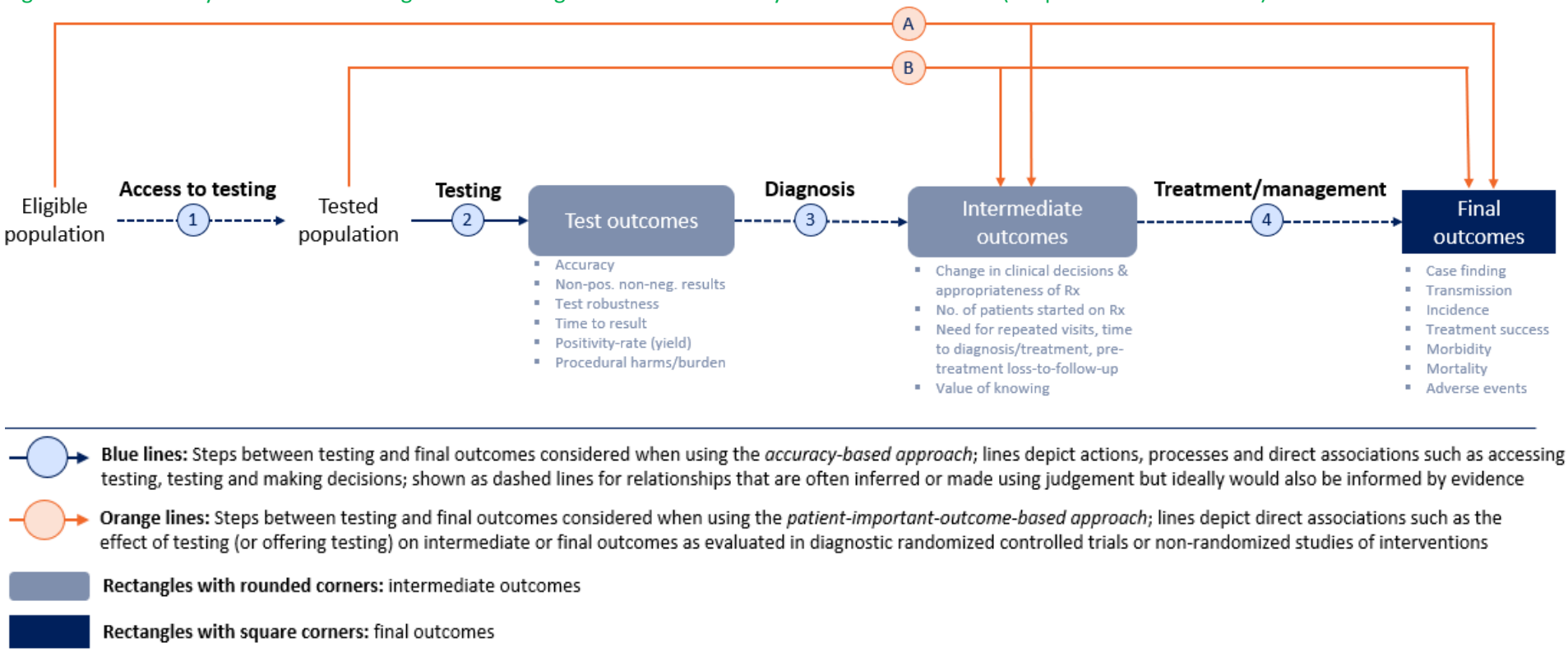
(B) Does the diagnostic intervention improve intermediate or final outcomes among the tested population?

A key difference between (A) and (B) is that the target population is defined differently and thus that the denominator for outcome measures differs.

Table 4.1 provides examples and further details on how users may link the information from key questions in the framework to relevant outcomes, sources of primary evidence, and considerations for evidence generation. Examples are not all-inclusive, and evidence generators are encouraged to critically evaluate their individual frameworks to identify further needs and opportunities for sourcing evidence and designing research evaluations.

[13] Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the U.S. Preventive Services Task Force. American Journal of Preventive Medicine. 2001 Apr;20(3):21–35.

Fig 4.1: Generic analytical framework to guide evidence generation on accuracy and health outcomes (adapted from the USPSTF)



***Non-positive non-negative results**: Group of test and instrument-related outcomes occurrence of which leads to test results being neither positive, nor negative, including e.g. rates of instrument failures, test failures, invalid results or indeterminate results. **Test robustness**: A test's ability to remain unaffected by variations in environmental conditions (e.g. temperature, humidity, dust), employment by users with varying levels of training or experience, and differing levels of adherence to test procedures. **Value of knowing:** An intermediate outcome reflecting any consequence for the wellbeing of a patient, or their family members or carers, that arises directly through the knowledge or information obtained as a result of testing (e.g. the value of having a diagnosis confirmed by a test, even if clinical management is not affecting by the result), rather than as a consequence of changed clinical management and related effects on health outcomes. **Positivity-rate** (diagnostic yield or simply yield): The positivity-rate of a test is the proportion of people in whom it indicates presence of the target condition among all people to whom testing was offered.*

Table 4.1: Examples for linking steps in the analytical framework to outcome measures, primary sources of evidence and other important considerations

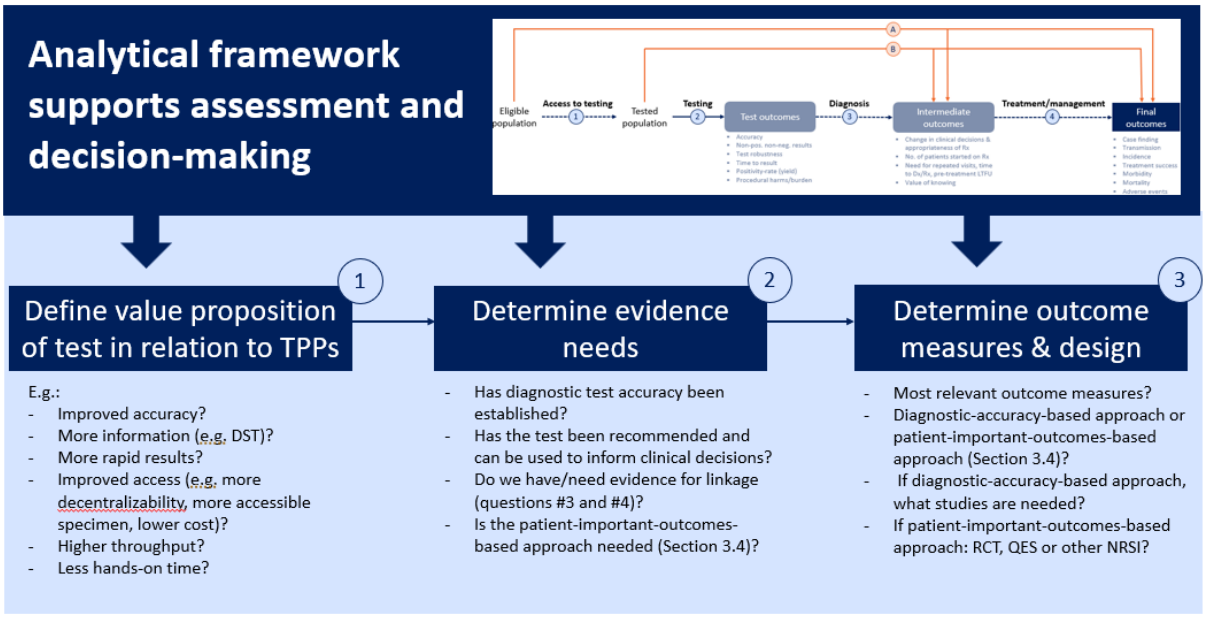
Step	Process	Outcomes related to process	Primary source of evidence	Important considerations
1	**Access to testing**	• Patient care seeking behaviour, e.g. % accessing health care and testing services • Ease of obtaining, storing or transporting specimens, e.g. % providing specimen • % that find the test acceptable?	• Cross-sectional studies • Patient pathway analyses • Care cascade analyses • Standardized patient studies	• Expected test placement & implementation, cost • Platform robustness, suitability for implementation at lower HS levels • evidence on whether decentralization leads to better access to testing
2	**Testing**	• Accuracy • Non-pos. non-neg. results • Test robustness • Time to result • Positivity-rate (yield) • Procedural harms/burden	• Cross-sectional studies (e.g. Diagnostic test accuracy)	
3	**Diagnosis**	• Evidence that changes in test outcomes lead to changes in intermediate outcomes (evidence supporting linkage)	• Non-randomized studies of interventions • Diagnostic before-after studies • Diagnostic randomized controlled trials and quasi-experimental studies	• Role of empiric therapy on decision-making • Relevance of ruling out TB for diagnosis of other conditions • Dependency on health system for test outcomes to affect intermediate outcomes
4	**Treatment management**	• Evidence that changes in intermediate outcomes lead to changes in final outcomes (evidence supporting linkage)	• Treatment trials • Evidence on natural history	• If patient spectrum detected by the index test is different from patient spectrum in which the net benefits of TB treatment is known, need to consider carefully • If spectrum is same, we know TB treatment works and benefits clearly outweigh the harms
A & B	**Direct effects on intermediate/final outcomes**	• Change in clinical decisions & appropriateness of Rx • No. of patients started on Rx • Need for repeated visits, time to Dx/Rx, pre-treatment LTFU • Case finding • Transmission • Incidence • Treatment success • Morbidity • Mortality • Adverse events	• Diagnostic randomized controlled trials • Quasi-experimental studies	

** Treatment success is declared in individuals who have been cured or completed treatment without recurrence (this implies absence of death, treatment failure and loss to follow-up during treatment and generally acquisition of drug resistance).*

Development of an analytical framework is therefore useful for determining the value proposition, evidence needs, study design, and outcome measures for TB diagnostic interventions. Figure 4.2 outlines key questions and considerations for each of these initial steps for TB diagnostic evidence generation. These steps are further detailed in sections below.

It is important to note that these sections focus on tuberculosis, however, the principles may apply to evidence generation for diagnostic interventions associated with other communicable or non-communicable diseases.

Fig. 4.2. Using an analytical framework to determine the evidence needs, outcomes and design of TB diagnostic evaluations



DST: Drug-susceptibility testing, RCT: Randomized Controlled Trial, QES: Quasi-experimental study, NRSI: Non-randomized studies of interventions

4.2. Developing a value proposition with an analytical framework

Key message 1 – Define the value proposition and prepare an analytical framework

Develop a description and visual representation of the clinical pathway in which the new TB test would be used, its role, how final patient-, population- or programme-level outcomes might be affected compared to standard of care and through which intermediate outcomes or mechanisms (i.e. an analytical framework; see Figure 4.1). Development should be based on existing WHO Target Product Profiles (TPPs), consultation with key stakeholders (e.g. affected community, testing personnel, other health care providers, policy makers etc.). Assess the evidence needs for each step in the pathway, for the linkages between steps, and for the affected outcomes. Determine what evidence is already available and sufficient to support demonstrating that the index test is likely to improve patient-important outcomes and plan evidence generation to fill important gaps.

Why this is important

Developing a clear understanding of how a new diagnostic test, specimen type, or testing strategy fits within the clinical pathway is a critical first step in evaluating its potential value. Visual tools—such as analytic frameworks, logic models, evidence models, causal pathways, decision trees or clinical flow

diagrams—can help clarify where and how a test is intended to be used, its relationship to the target population and comparator tests, and its expected impact on patient-important outcomes (see also section 4.1 and the intended purpose section of diagnostic TPPs). This can guide selection of one or more appropriate study design(s), evidence capture and reporting needs, and support alignment within study teams and between stakeholder groups. A detailed description of individual actions/processes and outcomes in the analytical framework, including a description of what evidence will be used to support each step or linkage should be prepared. Involving multiple stakeholders during development ensures a complete and balanced understanding of evidence needs.

4.3. Generating evidence on diagnostic accuracy

Generating robust evidence on diagnostic accuracy is a critical step towards any initial recommendation of a new TB test, specimen type, or testing strategy and a core component of evidence needed within the diagnostic accuracy-based approach to developing policy recommendations on diagnostic interventions. Generating evidence on diagnostic accuracy is typically the most time- and resource-intensive component of evidence to generate compared to other evidence used for initial policy recommendations (i.e., acceptability, feasibility, equity, and resource requirements).

Therefore, this section describes key messages to consider across major study protocol elements of diagnostic accuracy studies. Studies intended to support regulatory submissions should align with internationally recognized standards such as ISO 15189 and CLSI guidelines.

4.3.1. Study design

Key message 2: Design studies to minimize risk of bias

Design: For studies of tests for TB detection, use a cross-sectional or cohort study design and avoid case-control designs. Either a consecutive series or a random sample of people who require evaluation for TB should be enrolled. Prospective studies and testing of fresh specimens is preferred and should make up the majority of evidence; retrospective studies are also acceptable if design principles described in this guidance are followed and the impact of storing specimens is well understood. For tests of drug resistance, some enrichment for patients with increased risk of resistance may be used (see also section 4.3.2 on participant selection criteria). Comparative diagnostic accuracy studies, providing direct head-to-head comparison of the diagnostic accuracy (as assessed against the reference standard) of the index test to relevant standard of care tests are generally preferred (see sections 4.3.10 and Annex A3.1).

Blinding: If any subjectivity is involved in interpreting index test results, readers should be blinded to any other information about the participants, especially results of other tests, including the reference standard. Classification of patients with a composite or clinical reference standard (CRS) must be independent of the results of the index test and thus blinding should also be implemented as appropriate.

Timing*: Any time difference between collecting specimens for the index test, reference standard, and comparator tests should be minimized. For TB detection, a difference of a few days is generally acceptable, provided no treatment is initiated between sampling time points, as treatment could alter the disease state and affect test results. For tests to detect drug resistance, including people on treatment but not responding well as part of the population is acceptable. Considerations may vary for other tests or indications.

Follow-up: Follow up may be considered, in particular where there are concerns about the accuracy of the reference standard (see section 4.3.9 on reference standards), e.g. if the reference standard has low sensitivity and the index test has potential to exceed it (or do so in a subset of patients), follow-up of untreated study participants could capture later disease onset.

* Note: The critical need to diagnose TB rapidly and initiate treatment promptly once a diagnosis is made, according to existing guidelines, should never be undermined by requirements for certain sampling procedures of research studies.

Why this is important

Study design needs to be chosen to support the specific and relevant research question with respect to intended use of setting, target population and outcome measures of accuracy. The specific categories described above are considered by systematic reviewers and WHO when judging the risk of bias in included studies (based on QUADAS-2), together with other considerations in the following sections. As such, design choices have important implications for determinations on the certainty of evidence and strength of recommendations included in WHO policies (see section 3.3).

4.3.2. Participant selection criteria

Key message 3: Align selection criteria with the target population

Inclusion criteria should be clearly defined and ideally aligned with the intended target population and role of the test, such that an appropriate and representative patient spectrum is included. Populations with a lower bacillary load or who cannot expectorate sputum, such as people living with HIV and children, should be included early in test evaluations. Individuals already diagnosed with TB or started on treatment should be excluded from studies evaluating tests for TB detection. If the index test can be done on non-sputum specimens, include people unable to provide spontaneously produced sputum specimens (with reference standard testing done on induced sputa). If study participants can provide a specimen for one test but not for another—for example, if a tongue swab or urine specimen is available but sputum cannot be produced spontaneously —this difference in specimen availability may reflect inherent test characteristics. Therefore, it should not be considered a reason to exclude the participant or specimen from the analysis. Therefore, it should not be considered a reason to exclude the participant or specimen from the analysis.

For tests of drug resistance, patients with increased risk of resistance, or even patients already on treatment but not improving, can be considered for inclusion. However, this may skew the patient spectrum and needs to be considered during analysis, unless it is reflective of the target population of the test (see section 4.3.11 on analysis and reporting).

Why this is important

Careful consideration of eligibility criteria enhances the relevance of study findings and the populations to which they can be applied, supporting efforts to promote health equity. A representative patient spectrum (in terms of bacillary burden or other key drivers of index test sensitivity and specificity) is a critical requirement for reliable accuracy estimates. The population selected and ideally well-represented in terms of sample size for participation in clinical studies of a novel TB diagnostic intervention is a key factor in determining the generalizability of results. Including populations for whom TB diagnosis is more challenging, such as people living with HIV and children, is essential to avoid perpetuating evidence gaps and health inequities. Excluding such groups risks delaying or denying the benefits of innovation. In certain cases, extrapolation of findings may be possible but will not be sufficient for WHO to issue strong recommendations, as described in Annex 3. Including people unable to provide spontaneously produced sputum specimens is critical for tests that

can be performed on non-sputum specimens to enable evaluation of their added value for the diagnosis in these populations.

4.3.3. Study setting – participant enrolment

Key message 4 – Carefully consider selection of settings for participant enrolment

Ideally, the setting for participant enrolment is aligned with the eventual target setting(s) of use, considering geography, disease prevalence, and level(s) of the health system. Tests should be evaluated in a variety of geographic regions to capture strain diversity, heterogeneity in TB epidemiology, drug resistance mutation patterns and other factors that may affect test accuracy.

Why this is important

Test performance may vary by clinical setting due to people presenting with different stages of the disease. For example, for TB detection, testing of patients recruited in tertiary care settings may lead to overestimation of accuracy as patients often present at later stages of disease. For tests of drug resistance particular care needs to be given to this choice, depending on the drugs to be tested for, given wide variation in drug resistance prevalence and resistance mechanisms across geographies and sites. Evaluations from a single study in a single setting limit evaluation of inconsistency of results, thereby impacting the certainty of the evidence for this domain in GRADE (see section 3.3). Further, broad and representative geographical coverage can facilitate extrapolation and uptake of new recommendations to all regions and countries.

4.3.4. Study setting – testing

Key message 5: Generate evidence on the index test in its intended setting of use

Ideally, the setting where specimen collection, processing and testing of specimens takes place is aligned with the eventual intended setting of use for the index test (and comparator tests, as relevant), whereas the reference standard should be carried out under the best possible conditions to ensure optimal quality and validity of results. If it is not feasible or scientifically appropriate to conduct the entire study in the intended setting of use, then at minimum, a portion of the study should still take place in the intended use setting.

Why this is important

Although, performing the diagnostic intervention in well-equipped reference laboratories with stable infrastructure (e.g., reliable electricity and connectivity), optimal environmental controls (e.g., temperature, humidity), highly trained personnel familiar with molecular techniques and best laboratory practices ensures rigorous evaluation of test performance, it may overestimate test performance vis a vis use in real-world settings and cannot meaningfully capture feasibility, usability, and acceptability. To address this gap, conducting at least part of a study in the intended setting of use can provide evidence on the generalizability of data from testing in controlled laboratory environments. Aligning the setting for testing with the eventual setting of intended use also can allow for gathering meaningful evidence on the feasibility, acceptability and possible limitations of testing procedures (see section 4.6). Such studies also help develop practical resources—like standard operating procedures, sampling protocols, training materials, and testing workflows—to support effective and rapid implementation of the new diagnostic test.

4.3.5. Specimen processing and testing

Key message 6: Carefully consider and describe specimen processing and testing

Clearly describe the specimen flow within the study, including detailed procedures for specimen collection, handling, and processing for each test performed.

Specimen processing and testing: Specimens should be processed and tested as per the manufacturers' instructions (if available) or standardized final protocols for index test, comparator and reference standard. While adaptive study designs may have a role in optimization studies, procedures should not be changed during conduct of studies intended to inform policy development. If software is involved in generating test results, the same version should be used throughout a study. Similarly, if a testing process requires multiple procedures and instruments (i.e., targeted next generation sequencing end-to-end solutions) only those validated by the manufacturer for use with their technology should be used.

Specimen flow & head-to-head comparisons: Ideally, when evaluating new diagnostic tests, one specimen or multiple same day specimens collected should be used for index test, comparator and reference standard. A design permitting direct head-to-head comparison to relevant comparator tests is highly desirable, but it is critical to ensure giving equal opportunity (for e.g. randomizing the specimens collected on same or different days to various testing procedures) in terms of getting high-quality specimens of sufficient volumes to index test, comparator and reference standard *[Note: we may want to elaborate more on this and give example specimen flows in an Annex].* When evaluating new specimen types, consideration should be given to the order and timing of specimen collection according to specimen-specific optimization studies (for e.g. it is recommended to collect tongue swab before sputum collection or wait for 30 mins after sputum collection).

Banked specimens: Banked specimens may be used for part of a study if processed and stored appropriately according to manufacturer Instructions for Use and if convincing evidence (e.g equivalency studies) can be provided to demonstrate that storage does not affect the performance of the index test. Use of banked specimens is more acceptable for resistance testing than TB testing, based on drug resistance prevalence. Derived specimens (i.e., specimen matrix spiked with TB culture or DNA) may be applicable for WHO prequalification or regulatory assessments for limit of detection or reproducibility studies but are not included in diagnostic accuracy evaluations used for WHO guidelines as insufficiently direct for quantification of clinically relevant sensitivity and specificity values and ranges.

Why this is important

A clear description of the specimen flow and testing process is essential to permit interpretation of the evidence from any study, in particular for studies on TB diagnostics. Many specimen types for TB tests are limited in volume and often very heterogenous and as a result, how specimens are treated, split or attributed to different tests (index test, comparator, reference standard) can have important effects on performance estimates.

4.3.6. Extrapolation to excluded populations, settings, or specimen types

Key message 7 – Formulate a strategy for extrapolation of evidence to excluded populations, settings or specimen types

It is preferable to have sufficient data of all relevant populations, settings and specimen types such that conclusions can be made on direct and precise estimates of diagnostic accuracy; however, extrapolation or partial extrapolation to some excluded or imperfectly represented settings,

populations or specimen types can be justifiable in some cases if it is supported by relevant evidence. Generation of evidence that could support extrapolation --such as including at least a small number of participants representing an otherwise excluded subgroup, conducting part of a study in specific settings, or relevant bridging or equivalency studies-- should be considered when planning a study.

Why this is important

Generating evidence on all relevant populations, settings and specimen types is typically not possible in single studies and e.g. obtaining sufficient data on children and for people with extrapulmonary tuberculosis is often challenging. Especially during an early assessment of a test and extrapolation can sometimes be considered. Before extrapolating the evidence to other populations, it is important to consider disease prevalence (as it will impact test's predictive values) and spectrum of the disease in those specific populations and relevance of specimen matrix effects. Extrapolation to populations characterized by low bacillary burden can sometimes be supported by sensitivity estimates stratified by measures of bacillary burden (see section 4.3.11 on analysis and reporting). Extrapolation to closely related specimen types may be possible where there is evidence to support this (for example, when there is evidence that bacillary burden is similar in different specimens, there are no differences in matrix effect or interfering substances expected, or relevant and reliable analytical study data are available[14]).

4.3.7. Sample size

Key message 8 – Ensure sufficient sample size to achieve precise estimates

When determining the sample size of a study, consider levels of desired precision for sensitivity and specificity estimates, using the Wilson score method[15,16]. In general, a body of evidence comprising ~300 participants with confirmed TB or more (as defined by the reference standard) including approximately 30% of paucibacillary specimens yields reasonably precise estimates of index test sensitivity, with further returns in precision in estimates typically diminishing sharply beyond this number (see annex 4.2). Key subgroups, as per the value proposition of the test, should be adequately represented. Since typically evidence is synthesized through systematic reviews and meta-analysis, a careful assessment of evidence already available (and ideally what is known to be in progress) should be conducted before mounting new studies to optimally judge what is needed or appropriate. Where relevant, researchers may reference WHO guidelines, or guidance provided by WHO PQ and national regulators for sample size thresholds used to establish existing classes of TB diagnostics and historic examples of included studies and related meta-analyses that highlight successes and gaps in sample sizes for overall populations and subgroups of interest.

When planning comparative diagnostic accuracy studies, the paired nature of the data that will be generated should be taken into consideration when planning study size (e.g. by using Tango's score interval and justifiable levels of correlation during simulations).

[14] Food and Drug Administration. Leveraging existing clinical data for extrapolation to pediatric uses of medical devices: Guidance for industry and Food and Drug Administration staff. Silver Spring (MD): U.S. Dept. of Health and Human Services, Food and Drug Administration; 2016 Jun 21.

[15] Newcombe R.G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine.* 1998;17:857-872.

[16] U.S. Food & Drug Administration. Reporting Results from Studies Evaluating Diagnostic Tests: Guidance for Industry. 2007.

For resistance testing, sample size planning should consider levels of desired precision for sensitivity and specificity estimates separately for each drug being evaluated.

See also Annex A4.2

Why this is important

Small sample sizes result in wide confidence intervals or missed clinically important differences that could impact EtD judgements, certainty of evidence (downgrading for imprecision, see Section 3.3 and Annex 3). Careful planning of sample size, considering already available evidence, is therefore an important step during study design. Sample size calculations should also consider expected subgroup heterogeneity and allow for stratified analyses to detect spectrum effects to avoid downstream exclusion of important subgroups from recommendations due to lack of sufficient evidence.

Sample size requirements vary depending on the main objective and design of the study. For instance, equivalence or non-inferiority studies—which aim to demonstrate that a new test performs similarly to or not worse than an existing standard of care within a pre-specified margin—may require fewer specimens than studies designed to estimate diagnostic accuracy.

BOX. Index test, reference standard test, and comparator tests: What are the differences?

[Note: see these and other definitions also appear in the Glossary on page x]

Index test: The test under evaluation, sometimes just referred to as "test" or "diagnostic" in this document. We mainly use this term when the test is evaluated for its diagnostic accuracy in detecting a target condition, i.e. in the context of a diagnostic test accuracy study where test results typically are not used to inform clinical decision-making.

Reference standard: Is the test (or combination of tests) used to classify patients as having or not having the target condition. The reference standard is a measurement tool used to define sensitivity and specificity, not necessarily a test to compare the index test to (unless the reference standard is standard of care) and thus is often distinct from a relevant comparator test.

Comparator (or comparator test): A comparator test or strategy is the test or strategy reflective of current standard of care for routine clinical use in a given setting and/or the recommended test for use (based on international or local policy). In some instances, it may be identical to the reference standard test but often it is not, e.g. liquid culture is used as a reference standard but is not standard of care for TB detection in high-burden countries. In the context of a diagnostic test accuracy study, the purpose of a comparator is akin to the purpose of a control group in a randomized controlled trial as it permits direct comparison of outcomes between the new intervention and standard of care. In the context of a Diagnostic randomized controlled trial, the comparator is used to guide management in the control group.

4.3.8. Index test

Key message 9 – Provide a clear and comprehensive description of how the index test was applied

Provide a clear description of how the index test was applied in the study, including a description of training provided, number and type of retraining's that were required, and context/setting of where and how testing was done. If the test readout is not automated and requires a degree of subjective

interpretation, pre-specification of cutoffs for positivity and blinding of readers to other test results are essential, and inter-reader reliability needs to be assessed. For any inconclusive test results or test failures, it is important to provide details if repeat testing was done on those specimens and how they were analysed. For testing strategies, such as concurrent testing or pooling of specimens, or other complex interventions, provide clear descriptions about the sequence of steps, specimen management, testing and procedures and any intervention components that go beyond the use of a test. For new specimen types, provide a detailed description on how specimens were obtained, in particular if it not based on already widely implemented methodology.

Why this is important

Providing a clear and comprehensive specification of each component of the intervention is critical for informed judgements across EtD criteria, the certainty of evidence and implementation considerations. Understanding how an intervention was delivered in a study is critical for evaluating its feasibility under programmatic conditions and for anticipating how deviations from the study setting—for example, differences in specimen handling or procedural management—might affect diagnostic accuracy or outcomes. This will help assess the applicability and generalizability of the study outcomes for policy recommendations.

4.3.9. Reference standard

Key message 10 – Select an appropriate reference standard

[Note: we may provide more information and examples in an Annex. Some additional content for consideration is in Annex 4.1 already]

Ideally, all individuals who receive a novel diagnostic intervention should be tested with the same, highly sensitive and specific reference standard. A **microbiological reference standard** based on multiple liquid cultures (ideally from multiple specimens obtained on different days) is often the preferred reference standard for primary analyses when assessing tests to detect pulmonary TB. A single liquid culture may also be considered; however, a solid culture alone should not be used as a reference standard. Close attention must be paid to appropriate specimen processing and transport to avoid high contamination rates and overall quality control, and quality assurance best practices should be followed to ensure results are accurate, timely, and reliable. Positive cultures must undergo testing for confirmation of the presence of *M tuberculosis* complex bacteria. Regular external and internal quality control measures should be taken, and these results should be documented. Molecular assays should not be used as part of the microbiological reference standard.

For detection of pulmonary TB, if sputum cannot be produced spontaneously, induced sputum is a preferred specimen type for reference standard testing. This applies to both sputum and non-sputum-based diagnostic intervention. For extrapulmonary specimens, relevant specimen types based on the form of TB (e.g. cerebrospinal fluid, pleural fluid, lymph nodes etc.) should be used for testing with the reference standard.

A **clinical or composite reference standard** may also be considered to supplement analyses based on the microbiological reference standard where the microbiological reference standard lacks sensitivity (e.g. for paediatric and extrapulmonary TB). The components included in a composite reference standard need careful consideration and justification as each choice has implications for possible inconvenience for participants, resources and risk of bias (with trade-offs being inevitable

[17]). Ideally, a composite reference standard should be clearly defined, with its components standardized and applied consistently across all study participants. However, this may not always be feasible or ethically appropriate. It is also essential to specify which components, when positive, would determine the overall classification of the composite reference standard as positive. Components of a composite reference standard (beyond liquid culture on respiratory specimens) may include additional microbiological tests, non-microbiological tests (i.e., chest x-ray), and clinicians' decision to start TB treatment[18]. In extrapulmonary TB, ancillary tests such as histopathology, cytology, imaging, and body fluid analysis are important for a composite reference standard. For participants who are negative by the microbiological reference standard and not started on empiric treatment at enrolment, researchers should consider repeated reference standard testing, if feasible and clinical follow up for symptom resolution within the following two months for a more accurate classification.

Advanced statistical techniques such as latent class modelling may also be considered as they can account for the imperfect nature of microbiological reference standard and incorporate information from other tests or factors affecting pre-test probability.[1920]

Drug resistance testing: Phenotypic DST remains the reference standard for drug resistance testing of all anti-TB drugs, except rifampicin, pyrazinamide and ethambutol, where a composite reference standard combining phenotypic DST with whole-genome sequencing should be used. Research on whether a composite reference standard should also be used for bedaquiline is ongoing[21]. When sequencing is used, evidence generators should consider that not all mutations associated with resistance are known, and some mutations identified might not be associated with resistance. The latest WHO mutations catalogue should be referenced to ascertain relevant mutations and their associations with resistance. For drugs where WHO recommendations are not yet available, published and validated research methods may be considered.

See further detail in Annex A4.1

Why this is important

The choice of the reference standard has large implications for the meaning, interpretability and possible risk of bias of estimates of sensitivity and specificity. Careful choice of the reference standard should aim to avoid known biases, such as misclassification bias, differential verification bias, partial verification bias, incorporation bias, review bias, bias due to a composite reference standard and bias when comparing tests using non-comparative studies[16].

4.3.10. Comparators

Key message 11 – Include a comparator in the study

Include at least one relevant comparator for all study participants that represents the standard of care in the study settings. For TB detection tests, relevant comparators are WHO-recommended molecular

[17] Dendukuri N, Schiller I, de Groot J, Libman M, Moons K, Reitsma J, van Smeden M. Concerns about composite reference standards in diagnostic research. BMJ. 2018;360:j5779. doi:10.1136/bmj.j5779

[18] Graham SM, Cuevas L, Jean-Philippe P, Browning R, Casenghi M, Detjen AK, et al. Clinical case definitions for classification of intrathoracic tuberculosis in children: An update. Clin Infect Dis. 2015;61(Suppl 3):S179-S187.

[19] Rutjes et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment. 2007

[20] Schumacher, et al. Diagnostic Test Accuracy in Childhood Pulmonary Tuberculosis: A Bayesian Latent Class Analysis. American journal of epidemiology. 2016 Nov

[21] Köser CU, et al. A composite reference standard is needed for bedaquiline antimicrobial susceptibility testing for Mycobacterium tuberculosis complex. *J Clin Microbiol.* 2024;62(4)

diagnostics and smear microscopy as these represent the global or national standard of care for most populations. For new specimen types, relevant comparators are WHO-recommended specimen types that represent standard of care for the target condition (i.e., sputum for pulmonary TB) using the same or a standard-of-care test type. For certain population groups, additional tests may be considered as comparator tests or part of the comparator testing strategy (e.g. LF-LAM in people living with HIV or testing of stool in paediatric populations). For drug resistance tests, comparators may include genotypic or phenotypic standard of care methods, depending on the drug or drugs being targeted by the index test. A clinical decision to treat (empiric therapy) is also part of standard of care and thus may be considered in secondary analyses as part of a comparator (and index test).

Why this is important

In addition to establishing the diagnostic accuracy of a diagnostic intervention, specimen type, or testing strategy using an appropriate reference standard, it is important to compare its accuracy directly against the standard of care. First, this provides direct evidence for assessing whether a new test, specimen type, or strategy performs better, worse or similar to what is currently recommended or in use. Secondly, diagnostic test accuracy is not a fixed properly of a test but depends on the patient spectrum in a study or set of studies; a comparator with well-established performance acts as a sort of calibrator, leading to increased interpretability of the data because comparative diagnostic test accuracy is less dependent on the patient spectrum.[22] Lastly, data on the standard of care permits the assessment of incremental value if the new intervention is being considered as an add-on test.

4.3.11. Analysis and reporting

Key message 12 – Report transparently and provide comprehensive analyses

Preparing data for analysis: Investigators submitting unpublished data may be requested to format their data according to the research PICO questions, including disaggregation of findings by relevant subpopulations (i.e., children or people living with HIV) and sampling methods (i.e., healthcare-worker or self-collected specimens). *[Note: we plan to provide a sample set of indicators in an Annex which should be considered when submitting data for WHO guidelines.]*

General reporting: To ensure transparent and complete reporting of diagnostic accuracy studies, the Standards for Reporting Diagnostic Accuracy (STARD) checklist[23] should be followed. Provide a schematic diagram outlining when specimens are taken, what volumes were used, any splitting, what tests performed etc. (see section 4.3.5 on specimen processing and testing). Report on setting of participant enrolment, case-finding strategy (passive vs active case finding) and setting where testing with index and comparator tests took place. To facilitate assessment of the patient spectrum in a study, provide descriptive statistics on TB prevalence and measures of bacillary burden in clinical specimens. Provide sufficiently detailed description on methods used for all tests (index test, reference standard and comparator(s) as well as observed culture contamination rates.

Comparative accuracy estimates: When comparative evidence is available, report on comparative accuracy (i.e. the difference in sensitivity and specificity, with 95% confidence intervals around these differences), in addition to estimates of diagnostic accuracy of the diagnostic intervention itself. For paired data, use appropriate methods for computation of confidence intervals (e.g. Tango's score

[22] Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy. Ann Intern Med. 2013 Apr 2;158(7):544–54.

[23] Bossuyt P M, Reitsma J B, Bruns D E, Gatsonis C A, Glasziou P P, Irwig L et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies *BMJ* 2015; 351 :h5527 doi:10.1136/bmj.h5527

interval). If participants can provide a specimen for one test but not another (e.g. able to provide urine but not sputum), several analyses should be performed and reported: (i) an analysis restricted to participants/specimens where results are available for both tests and (ii) analyses on the remainder, highlighting differences in robustness and yield, ideally with reference standard results also available in this group.

Stratification: Provide 2x2 tables and stratified accuracy estimates for relevant subgroups, e.g. people living with HIV, children, and a measure of bacillary burden (for example by smear-status, smear-grading or semi-quantitative results of molecular tests). For tests of drug resistance, where patients are already on treatment but not improving can be considered for inclusion, stratified analyses for patients tested prior to versus during treatment should be provided.

Discordant (index test-positive, culture-negative) results: Provide line listings of discordant test results for any detailed analyses on these specimens. Understanding discordant results is important, but additional investigation of discordant results or relating specimens cannot be used to change estimates of diagnostic accuracy. To understand discordant results consider the following, as applicable depending on the nature of the index test: (1) in silico analyses and exclusivity studies before study initiation *[NOTE: will explain in definitions or annex]*; (2) following-up patients to uncover subsequent culture conversion and examination of alternative diagnoses; (3) environmental testing during the study to assess potential for specimen management challenges or cross-contamination; (4) sequencing of amplicons to detect potential nonspecific amplification or presence of mutations not detected by an index resistance test; (5) rigorous assessment of prior treatment for TB; (6) exploration of other patient- and setting-specific characteristics that may lead to false-positive results and (7) conducting additional analyses using tests beyond the reference standard, a composite reference standard or both.

Estimands: Clear specification of estimands may enhance alignment between study goals, methods used for analysis and interpretation of results. Alternative estimands may be considered for sensitivity analyses, where applicable.[24,25] *[See also Annex A4.4.]*

Why this is important

Preparing data for analysis: Systematic reviews to support policy development are usually based on published scientific articles. Typically, patients are used as the unit of analysis; if specimens are used, the clustered nature of the data must be appropriately considered. If study data have not been published, but a quality data set has been completed, cleaned and locked, such unpublished data can be submitted to WHO. Published data and locked unpublished data will be assessed according to the research questions for each guideline development process.

General reporting: Complete reporting in line with international guidelines is critical to allow for complete assessment and understanding of the evidence. Transparent reporting on financial disclosures such as involvement of the manufacturer in the study is critical to assess any potential conflict of interest which could impact the study results.

Comparative accuracy estimates: see section 4.3.10

Stratification: Meta-analyses typically summarize data as presented in the published literature and it is therefore highly desirable that relevant 2x2 tables and stratified analyses –even if numbers of an individual study are too small to generate reliable estimates on their own-- are presented in

[24] Fierenz A, Akacha M, Benda N, Badpa M, M M Bossuyt P, Dendukuri N, et al. The Estimand Framework in Diagnostic Accuracy Studies. Statistics in Medicine. 2025;44(20–22):e70248.

[25] Evans SR, Pennello G, Zhang S, Li Y, Wang Y, Cao Q, et al. Intention-to-diagnose and distinct research foci in diagnostic accuracy studies. The Lancet Infectious Diseases. 2025 Aug;25(8):e472–81.

diagnostic accuracy studies. This permits extraction of all details required without having to contact authors to request this additional detail.

4.3.12. Data sharing

Key message 13 – Share individual participant data

De-identified individual participant and test data should be made widely available, preferably through established data repositories where data can be found and obtained through secure and standardized processes.

Why this is important

Making de-identified individual participant data publicly available provides the possibility for individual participant meta-analyses and other research, to gain further insights and understanding of the index test. Providing open and equitable but secure access to study data offers the greatest opportunity for learning and is in the spirit of open data. The use of existing data repositories can facilitate good data sharing practises (FAIR principles)[26], dissemination and access for further research on existing data. *[NOTE: we are not well aware of which platforms accept data on diagnostics; Vivli is one although it appears probably >90% of data is on trials)]*

4.4. Generating additional evidence as part of diagnostic accuracy studies and to complement the diagnostic-accuracy-based approach

Even when the effects of a diagnostic intervention are estimated primarily through a diagnostic-accuracy based approach, it remains essential to consider evidence beyond diagnostic accuracy. This includes additional outcomes that can often be assessed alongside accuracy studies such as time to result. Such complementary evidence should be evaluated together with diagnostic accuracy data (as described under desirable and undesirable effects section in the EtD, see section 3.4). These additional data can substantially influence the "balance of effects", which is a key determinant of direction and strength of WHO recommendations (see also Annex A3.6).

4.4.1. Non-positive non-negative results and test robustness

Key message 14: Provide a careful analysis of non-positive non-negative results and an assessment of test robustness

Non-positive non-negative results: Report borderline, unsuccessful (errors, invalid etc.) and missing results or instrument failures for index test, reference standard and comparator tests[27], as well as results from repeat-testing following such errors. Record unsuccessful test results of the reference standard and report test results from index and comparator tests among these.

Test robustness: Evaluate a test's ability to remain unaffected by variations in environmental conditions (e.g. temperature, humidity, dust), varying levels of training or experience, and differing levels of adherence to test procedures as appropriate based on the intended setting of use of the test.

Why this is important: The value of a highly accurate test will be limited if a test often fails to produce valid or interpretable test results. Non-positive non-negative results often necessitate

[26] FAIR Principles: https://www.go-fair.org/fair-principles/

[27] Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013 May 16;346:f2778. doi: 10.1136/bmj.f2778. PMID: 23682043.

repeat testing or recalling patients to provide a new specimen, creating additional burden and cost for both patients and health systems. Therefore, non-positive non-negative results are typically considered an important outcome during GDGs and may be considered when evaluating the desirable/undesirable effects of the diagnostic intervention in the EtD framework as well as considerations about the likelihood of test results affecting management.

4.4.2. Time to result

Key message 15: Measure time to result for the index test and comparator

Measure the time it takes from obtaining a specimen to getting a test result and compare it to relevant comparator(s). Provide estimates for time to result for different batch sizes or as a function of other relevant variables, depending on the test. If this is done in the context of a diagnostic accuracy study, measurement in a small subset of tests done is typically sufficient to yield reliable estimates.

Why this is important: Individuals accessing TB testing services and health care providers value rapid result availability [28, 29]. Accordingly, time to result is typically considered an important outcome during GDGs. Generating quantitative evidence on time to result permits inclusion of this important aspect as an outcome to be considered when evaluating the desirable/undesirable effects of the diagnostic intervention in the EtD framework. Tests with shorter time to result may also increase the likelihood of test results affecting management (another important consideration in the EtD framework).

4.4.3. Procedural harms / test burden

Key message 16: Evaluate possible procedural harms or burdens associated with testing

Evaluate the possible procedural harms or burdens involved for people tested in relation to the process for obtaining specimens or carrying out the test as compared to relevant comparator(s). This is particularly important if the effects on patients cannot be assumed to be equivalent to the comparator (e.g. different specimen type or differing direct interaction between those tested and the testing process).

Why this is important: Direct procedural burdens or other adverse effects or burdens of the overall testing process, including obtaining a specimen, are important and need to be considered as part of the judgement of desirable and undesirable effects. For TB tests this has often not been very important because in terms of any procedural harms processes were identical from the patient perspective. However, when processes differ more between diagnostic intervention and comparator (e.g. comparing tests that require phlebotomy vs tests done from sputum), this aspect may be more important.

[28] Engel N, Ochodo EA, Karanja PW, Schmidt B-M, Janssen R, Steingart KR, Oliver S. *Rapid molecular tests for tuberculosis and tuberculosis drug resistance: a qualitative evidence synthesis of recipient and provider views.* **Cochrane Database of Systematic Reviews.** 2022; Issue 4: CD014877. doi: 10.1002/14651858.CD014877.pub2.

[29] Shah K, Oswald L, Mabunda S, Karanja PW, Huddart S, Cattamanchi A, et al. *Preferences for tuberculosis diagnostic test features among people tested for tuberculosis: a multi-country discrete choice experiment.* **The Lancet Public Health and Respiratory Collection.** 2025; (in press).

4.4.4. Test-positivity rates (diagnostic yield)

Key message 17 – Consider conducting studies on test-positivity rates (diagnostic yield) if the diagnostic intervention may increase access to testing

Consider conducting studies on diagnostic yield (the proportion of people who test positive for tuberculosis among those to whom testing is offered) if the value proposition of the diagnostic intervention lies in providing greater access to testing, e.g. through the use of a more accessible specimen type. The use of diagnostic yield as an outcome measure may be appropriate after it has been demonstrated in previous studies that (i) the sensitivity and specificity of the intervention are acceptable according to WHO guidance (i.e., target product profiles or performance values for established classes of TB diagnostics), (ii) the specificity of the intervention is as good as that of the comparator, (iii) the specificity is not reduced in specific study settings or populations (including subpopulations), (iv) for tests of non-sputum specimens, the specificity among those who can and cannot produce spontaneous sputum is similar (as determined by use of sputum induction for testing with a microbiological reference standard), and (v) the net benefits of treatment for those who cannot be tested with a comparator or reference standard test are equivalent to those for whom net benefits are known. In studies of yield, it is essential to incorporate testing (or attempted testing) with a relevant comparator that represents standard of care in the intended setting and population of use (section 4.3.10). Carefully consider any differences in pre-test probability between the populations tested with diagnostic intervention and comparator. Comparisons in yield between a diagnostic intervention and comparator(s) may be done within-patients (i.e. both index and comparator done on all patients as possible, as often done in diagnostic accuracy studies) or by random allocation between patients (i.e. some patients receive the novel diagnostic intervention, others the comparator, as is done in diagnostic randomized controlled trials).

Why this is important

Equitable access to diagnostic testing is essential for achieving fair health outcomes across populations [REFs to UHC and WHO Dx Standard]. Disparities in access driven by socio-economic factors, geography, and other factors can lead to delayed diagnoses, under-treatment, and poorer health outcomes in marginalized groups. It is therefore important to generate evidence on how a diagnostic intervention influences access to testing. Diagnostic yield can serve as an indirect indicator of access by demonstrating that an intervention can: (i) expand availability of testing services (e.g. through greater decentralization), (ii) enable increased testing through the use of a more accessible specimen type, or (iii) allow testing at larger scale or among broader populations due to enhanced operational feasibility or lower cost. This dimension is not captured by diagnostic accuracy studies, which are typically limited to individuals already able to access existing testing services and provide specimens. Diagnostic yield can also be measured without a reference standard (e.g. liquid culture), enabling evidence generation in intended programmatic settings that better reflect real-world conditions faced by patients and providers. However, it is important to know the test accuracy and not simply rely on diagnostic yield. When interpreting such evidence, differences in pre-test probability between the diagnostic intervention and comparator should be carefully considered, as expanding access to populations with lower disease prevalence may result in higher absolute numbers of false positives, even if specificity is unchanged. Importantly, improved access to testing particularly for hard-to-reach populations may represent an acceptable trade-off for a modest reduction in test accuracy, provided the overall public health benefit and equity gains are substantial. Further considerations related to equity are discussed in Section 4.6.4.

4.4.5. Evaluation of multi-test diagnostic strategies

Key message 18 – Consider additional design and analytical aspects when evaluating diagnostic strategies comprised of more than one test

To evaluate diagnostic testing strategies that are based on more than just a single test, additional consideration with regard to design and analysis are needed. The combined accuracy of several tests should be evaluated based on studies where all tests were done each study participant, rather than by combining accuracy estimates on several tests from several different studies or each investigating only one test. The way results are combined (i.e. which decision rule is employed to lead to a final result) needs to be clearly defined.

If strategies go beyond simple combinations of a few qualitative tests (each leading to a "TB" or "not TB" output), appropriate statistical methodology should be employed for the selection of relevant variables or component tests and the incorporation of quantitative outputs.

If combining multiple tests leads to overall accuracy characterized by both imperfect sensitivity and specificity, e.g. in the case of scoring rules or multivariable prediction or diagnostic models, careful consideration needs to be given to a range of additional factors that are otherwise typically of little relevance to diagnostic evaluations,[30,31] e.g. what trade-offs may be appropriate to balance under-diagnosis and over-diagnosis[32] and related evaluation of clinical utility (eg, using decision curve analysis[33]).

Why this is important

Different test modalities may capture different aspects or forms of TB disease, i.e. they may not be perfectly correlated, and as a result, combining multiple tests into a diagnostic testing strategy can be useful in identifying more TB with two tests than with one. Since the degree of correlation cannot be reliably predicted by theory, understanding the value of combinations of tests requires evaluating them simultaneously in the same group of study participants. This is evident for example when considering the additional value of urinary LAM assays on top of molecular assays done on respiratory specimens, despite the much lower overall sensitivity of the LAM assay.

For more complex combination strategies --as e.g. used for treatment decision algorithms for paediatric TB, or which may become relevant with increased development of digital tools e.g. in the screening context-- several additional considerations become relevant. These include methods used for variable selection, use of quantitative outputs, consideration of trade-offs when deciding on a threshold and the metrics that may be used to judge performance, which are discussed in detail in the "prediction modelling" literature but typically receive little attention by evaluators of TB diagnostics as most of our tools are simple, single tests that lead to a binary "TB vs not TB" output.

[30] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350:g7594.
[31] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. 2024 Apr 16 [cited 2025 Sept 29]
[32] Pauker SG, Kassirer JP. The Threshold Approach to Clinical Decision Making. New England Journal of Medicine. 1980 May 15;302(20):1109–17.
[33] Vickers AJ, Calster BV, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ [Internet]. 2016 Jan 25 [cited 2025 Nov 6]

4.5. Generating evidence to support linkage across the analytical framework

Within the diagnostic-accuracy-based approach, evidence on test sensitivity and specificity, alongside other outcome measures outlined in Section 4.4, plays a central role. However, it is important to be aware that within WHO policy development important consideration is also given to what may happen in terms of diagnostic and treatment decisions, once tests results become available; or in other words to what degree we can expect improvements in test outcomes to affect intermediate and final outcomes (see Figure 4.1). Therefore, evidence that demonstrates linkages across components of the analytical framework can provide critical support to the guideline development process. Such evidence may, for instance, show that improved test outcomes lead to improved intermediate outcomes (e.g. that shorter time to result reduces pre-treatment loss to follow-up) or that a diagnostic strategy that enhancements in intermediate outcomes translate into improved final outcomes (e.g., that reduced pre-treatment loss to follow-up leads to lower mortality). In this chapter, we outline how such other types of evidence that support linkages across elements of the analytical framework can support WHO policy development.

Key message 19 – Generate evidence to link diagnostic accuracy data to changes in intermediate and final health outcomes

To complement evidence on diagnostic accuracy studies, consider generating (or synthesizing) evidence that contextualizes findings or helps to link steps in the chain of events that are displayed in the analytical framework. Relevant evidence includes studies demonstrating (i) how improving access improves testing levels, (ii) how better test outcomes lead to improved intermediate outcomes and (iii) how improved intermediate outcomes may lead to improved final outcomes. Combining multiple sources of evidence through mathematical modelling may be of value if assumptions are well supported by evidence.

Why this is important

The diagnostic accuracy–based approach to evaluating diagnostic interventions (see Section 4.4.3) draws upon multiple sources of data and outcomes spanning the steps of the analytical framework (see Figure 4.1). These elements must be appropriately linked by decision-makers involved in policy development to form a coherent understanding of the intervention's likely effects on health and outcomes. Confidence in how well results connect across these steps may vary and can be strengthened by additional evidence that contextualizes or substantiates these relationships. Such supporting evidence may include, data on the proportion of patients in a given setting who can access testing or provide specific specimen types; patient pathway analyses, care cascade analyses, or standardized patient studies; evidence on whether the placement of the intervention within the health system (e.g., decentralization) influences access to testing; data on the consequences for individuals unable to access testing or lost to follow-up; evidence examining whether turnaround time for test results is associated with time to diagnosis; and evidence linking time to diagnosis or treatment initiation with pre-treatment loss to follow-up or final health outcomes.

While such evidence does not provide direct outcome measures on its own, it can substantially strengthen the evidence base for policy-making, particularly when using a diagnostic accuracy–based approach, by clarifying and reinforcing the linkages between intermediate and final outcomes.

4.6. Generating evidence on patient-important outcomes

Once the diagnostic accuracy of a test or class of tests has been established, and an initial recommendation has been made for their use using the diagnostic-accuracy-based approach, they can be used to guide clinical decision-making. It is then sometimes useful or necessary to generate evidence that more directly addresses the question of whether the use of a new test or diagnostic intervention improves health and other outcomes by using the patient-outcome-based approach. The preferred approach to do this is by using diagnostic randomized controlled trials. While diagnostic randomized controlled trials are sometimes considered challenging or costly to conduct, they provide the most reliable evidence on the effects of a diagnostic intervention. Further, diagnostic randomized controlled trials are typically low risk trials – this means that it should be possible to conduct them as pragmatic trials, embedded into routine care, with randomization at the cluster level, limited data collection and reporting requirements beyond routine care.[34,35]

This Section outlines common situations where using the patient-outcome-based approach is particularly pertinent (Section 4.5.1) and which intermediate (4.5.2) and final (4.5.3) outcomes may be considered when it is taken.

4.6.1. Patient-important-outcomes-based approach

Key message 20 – Consider conducting diagnostic randomized controlled trials when judging tests' effects on health outcomes based on accuracy may not be reliable

Once a test is in clinical use, consider conducting diagnostic randomized controlled trials. Quasi-experimental studies may be considered as an alternative to diagnostic randomized controlled trials but are less preferred. Generating such evidence is most important when the accuracy-based approach to evaluating its health effects may not be reliable. Typically, this is the case when the diagnostic intervention (i) may lead to significant changes in program implementation (e.g. how and where testing is performed, like a change from centralized labs to point-of-care) or (ii) if a testing approach means that there are changes in the eligible population being tested (e.g. testing asymptomatic or lower-risk individuals, resulting in a possible change in the patient spectrum being diagnosed and treated). Using the patient-important-outcomes-based approach is also preferred when the best available reference standard is not accurate.

Why this is important

Certain tests are initially recommended largely by supportive evidence on their diagnostic accuracy, but also have broader potential to benefit patients and populations, e.g. because (i) they may provide more rapid results, thus potentially permitting treatment initiation within the same clinical encounter and reducing in pre-treatment loss-to follow-up etc.; or (ii) they may facilitate greater access to testing services e.g. by being more easily decentralizable, by facilitating increased testing through the use of a more accessible specimen type; or (iii) they may permit testing to be carried out at larger scale or of a wider population due to other reasons (like operational feasibility or cost).

If such benefits come at the cost of reduced diagnostic accuracy compared to the current standard of care, the accuracy-based approach becomes generally challenging and less reliable when trying to decide between alternatives. In this situation, the preferred approach is to generate comparative

[34] WHO. Guidance for best practices for clinical trials. 2024

[35] OECD Recommendation on the governance of clinical trials. OECD 2013

evidence on the effects of the new diagnostic intervention versus standard of care on (intermediate or final) patient-important outcomes directly, ideally using a diagnostic randomized controlled trial design.

If the best available reference standard is not accurate or cannot be properly applied (e.g. because the required specimen cannot be produced) the true disease status of individuals is uncertain. If tests under evaluation are imperfect compared to the reference standard (e.g. the reference standard is insensitive, but diagnostic intervention sensitivity is even lower), this typically still permits the use of the diagnostic- accuracy-based approach. However, if e.g. the reference standard is insensitive and the diagnostic intervention suggests presence of disease where the reference standard does not, it becomes challenging or impossible to judge the value of the novel diagnostic intervention or indeed to know what diagnostic approach is optimal. In this case the only reliable way of assessing the best approach to diagnosis may also be a diagnostic randomized controlled trial.

4.6.2. Intermediate outcomes

Key message 21 – Generate evidence on the effect of the diagnostic intervention on intermediate outcomes

Generate evidence on intermediate outcomes that may be affected by diagnostic interventions and which, in turn, are thought (or known) to affect final outcomes (see section 4.5.3). Intermediate outcomes include those related to clinical decision-making (e.g., change in patient management, appropriateness of treatment decisions, number of patients started on treatment); time (e.g. the need for repeat visits, time to diagnosis, time to treatment initiation, pre-treatment loss to follow-up); and the value of knowing (e.g., patient awareness of the result, the perceived importance of testing and emotional impact of results).

Various study designs may be used, depending on the intermediate outcomes, including diagnostic randomized-controlled trials, quasi-experimental studies and qualitative research studies.

Why this is important

Generating data on intermediate outcomes is important for two reasons: (1) improvements in these outcomes typically matters to patients directly (i.e., they are patient-important outcomes), and (2) some evidence suggests that changes in certain intermediate outcomes affect final outcomes (i.e., they are surrogates for, or predictors of, final outcomes such as case finding, treatment success rates, morbidity and mortality). Intermediate outcomes may also better capture effects of diagnostic interventions when implemented programmatically, where factors that are unrelated to index test performance often affect the degree to which diagnostic interventions can affect patient outcomes; this can point to important implementation considerations. Therefore, they complement evidence on diagnostic accuracy and other test outcomes (e.g., robustness, time to result), and are considered during guideline development. Intermediate outcomes may be considered during deliberations on the benefits and harms of an intervention and help to increase the linkage and credibility of indirect evidence from diagnostic accuracy studies with possible impacts on final outcomes.

4.6.3. Final outcomes

Key message 22 – Generate evidence on the effect of the diagnostic intervention on final outcomes

Where feasible and appropriate (see also section 3.6 and 4.5), generate evidence to support the patient-important-outcomes-based approach on the effect of introducing a new diagnostic intervention on final outcomes. Final outcomes may include patient-level health outcomes (e.g. treatment success, mortality, quality of life), and health systems or population-level outcomes (e.g. changes in TB case detection rates, diagnostic coverage, TB incidence and population-level mortality). Use diagnostic randomized-controlled trials to estimate effects of diagnostic interventions on final outcomes; typically, randomization at the cluster-level is most appropriate. When diagnostic randomized-controlled trials are not feasible, well-conducted quasi-experimental studies can also provide valuable evidence (see Annex 5 for more detail).

Why this is important

Final outcomes are typically prioritized by guideline development groups as critical outcomes, reflecting their importance to patients and programmes. Diagnostic randomized controlled trials of health outcomes and population-level outcomes often provide the best evidence to inform guideline development with randomization providing optimal protection against selection bias and confounding and the use of final outcomes reducing the uncertainties of linking multiple pieces of evidence together (as is done in the accuracy-based approach). However, conducting randomized controlled trials may sometimes not be feasible and other study designs that could provide evidence on final outcomes should be considered. Certain quasi-experimental study designs may offer a reasonable alternative to randomized trials as they may be less susceptible to biases that make interpretation of many other observational study designs challenging (see Annex 5 for more detail).

4.7. Generating evidence on values, cost, cost–effectiveness, equity, acceptability and feasibility

Deliberations on the remaining EtD criteria (values, resources required, cost–effectiveness, equity, acceptability and feasibility) should be informed by research evidence on these criteria because they are all pertinent for formulation of recommendations. Having evidence available on these criteria is particularly important if the evidence on desirable and undesirable effects suggests that, on balance, neither the diagnostic intervention nor the comparator test would be favoured (i.e. they appear to be equivalent in terms of benefits and harms).

Generating evidence on some of these EtD criteria can be relatively simple and low cost, compared to the cost of generating evidence on diagnostic accuracy or effects on health outcomes. If the evidence on desirable and undesirable effects suggests that the intervention and comparator are apparently equivalent in terms of benefits and harms, it may be possible to determine which test or testing strategy should be prioritized based on careful assessment of values, cost, cost–effectiveness, equity, acceptability and feasibility. Evidence on these criteria may be gathered alongside studies of diagnostic accuracy or as part of separate studies. Some general, overarching findings within and between WHO-defined classes of TB diagnostics are sometimes broadly applicable for many different guideline meetings or across the assessment of different tests.

4.7.1. Values

Key message 23 – Conduct research on values (i.e. the relative importance people place on health outcomes)

Values affect the weighting of desirable and undesirable effects, potentially modifying a recommendation based on the balance of effects derived from diagnostic test accuracy studies. Better data from quantitative and qualitative studies about how different stakeholders (especially patients but also testers, health care providers and policymakers) value different outcomes (e.g. false positives versus false negatives; access or time to result vs accuracy) would make it easier to incorporate this information into WHO TB diagnostic guideline development processes more explicitly. Various types of studies can be employed to this end, including studies estimating utility values directly, discrete choice experiments and qualitative research studies; such studies can be embedded in, or run independently of, studies of tests, and results synthesized in a systematic review. (see also Annex A4.2)

Why this is important

Judgements about the magnitude and balance of effects depend on the magnitude of effects but also on the importance of outcomes. For example, a small positive effect on an outcome that is of the highest importance to stakeholders may outweigh a larger negative effect on an outcome that is seen as less important. More direct evidence on the relative importance people affected by TB place on the different health outcomes considered during GDGs would be valuable. The strength of a recommendation can also be affected when there is uncertainty about how those affected by the relevant intervention value its outcomes.

4.7.2. Resources required

Key message 24 – Gather evidence on the resources required to deliver the diagnostic intervention

Data should be collected on testing costs from both health system, inclusive of implementation process costs, to generate estimates of the unit per-test cost. Unit test costs should ideally be estimated using a micro costing or bottom-up approach where possible. Component costs to consider include equipment, materials/consumables, human resources, overhead and transport. The total costs required for implementation process should be assessed on a per-site basis and for a unit cluster of the laboratory network. Data on implementation costs may include costs training, quality assurance, infrastructure, specimen transportation and feedback of results. Costs typically vary greatly between countries, healthcare settings, and by operational factors (e.g., service volumes, overheads); therefore, costing should be carried out across multiple countries, settings, and operational scenarios to allow for broad representativeness and utility of results.

Why this is important

For most new diagnostics, NTPs may struggle to speed-up the implementation and scale-up more expensive tests on a large scale, even if their use is “cost-effective”. Regardless of WHO recommendations, the cost of interventions significantly impacts their adoption. Likewise, the resources required for testing is an important factor and may affect the strength of recommendations; for example, some recommendations may be made ‘conditional’ when there lack data and evidence on availability of in-country resources and infrastructure or to inform types and level of resources required for initial implementation and scale up.

4.7.3. Cost–effectiveness

Key message 25 – Carry out cost–effectiveness analyses

Cost–effectiveness studies that combine estimates of resource use with estimates of health effects provide additional value to support decision-making and are critical in ensuring adopted interventions are likely to provide appropriate value for money. This is particularly the case if an intervention leads to improved health outcomes, but costs more than the current standard of care in a particular setting SOC. Evidence can be generated based on both systematic review of literature and use of decision analytic models. For model-based studies, it is important to consider the data and assumptions that inform the model to ensure it is representative of the research question, setting and patient population. CHEERS checklist should be used to ensure high quality reporting of economic studies. Uncertainties and variability in health effects, health system costs, resource use and willingness to pay thresholds all need to be carefully considered. Careful consideration of the comparator of interest is also critical as choice of comparator will lead to different conclusions around cost-effectiveness. Whenever possible, comparisons should be made with existing standard of care in settings of interest to not over or underestimate the potential impact of novel approaches.

Why this is important

Cost–effectiveness analysis can further inform GDG decisions by evaluating the incremental costs of a new TB test (against an SOC) per incremental health improvement. In situations where costs of novel tests or screening algorithms will clearly exceed those of the SOC, formal cost–effectiveness analysis can have an important influence on the strength of the recommendation. In instances where costs of novel tests or screening algorithms are lower than that of the SOC, it is important to assess both the degree of cost-savings and the direction of the effectiveness.

4.7.4. Equity

Key message 26 – Investigate the impact on health equity

Equity should be investigated using five steps[36]: 1) identify populations who may experience inequities[37], 2) determine baseline risk for prioritized outcomes in these populations, 3) evaluate the representation of these populations in the studies providing the evidence, 4) conduct subgroup analyses for these populations if possible and 5) identify barriers to implementation of effective interventions within populations experiencing inequities. Based on the findings from this investigation, it is useful to then quantify current health inequities and evaluate how the introduction of the new test may affect current health inequities or introduce new ones.

Why this is important

Guidelines can play a crucial role in promoting health equity by explicitly considering how recommendations affect populations at high risk and for people in vulnerable situations. This requires explicit consideration of whether and how the introduction of a novel test may improve or worsen existing health inequities or lead to new ones. For example, less complex tests that could be implemented widely and are accessible to all populations (including remote, underserved or other vulnerable groups) are typically more likely to increase equity, although effects may differ between population groups (e.g. increasing equity for a particular gender but decreasing it for others).

[36] Dewidar et al, JCE 2024

[37] World Health Organization. Tuberculosis among populations at high risk and people in vulnerable situations: policy brief. Geneva: World Health Organization; 2025.

Additional complexity could include a need for changes to patient pathways, DST, quality assurance and supply chains or storage.

4.7.5. Acceptability

Key message 27 – Investigate the acceptability of the diagnostic intervention

Cultural norms and the characteristics of a test affect its acceptability to patients, providers and policymakers. Such characteristics include the, overall ease of use, costs incurred to the patient, easy access to get tested, maintenance and calibration requirements and support systems, reagent kit storage/stability, specimen preparation steps, quality of training materials, connectivity, infrastructure requirements, specimen transport requirements, biosafety considerations, availability of other assays to use on the same instrument (for multi-disease testing), and an instrument's physical footprint as well as health outcomes expected to arise from its use. Acceptability should be measured directly through quantitative studies (e.g. stakeholder surveys, discrete choice experiments, best-worst scaling) or qualitative research studies; it may also be reflected indirectly through data on uptake of novel tests. Ideally, comparative evidence is generated directly within the context of a study, but generating more indirect evidence outside a specific study context can also be of value (e.g. by studying the acceptability of different real or hypothetical test attributes, or use of the same test for another disease condition).

Why this is important

Acceptability is a multifaceted concept that has been defined as "the extent to which people delivering or receiving a healthcare intervention consider it to be appropriate, based on anticipated or experienced cognitive and emotional responses to the intervention" [38]. Where a test is not acceptable to policymakers it will not be taken up by NTPs; if it is not acceptable to health care providers, they will hesitate to use it if they have alternative choices; and if it is not acceptable to patients, they may avoid testing or not trust test results. Therefore, when developing novel interventions, it is critical to think about what is likely to be acceptable to key stakeholders. Quantitative studies can provide information on the acceptability (e.g. the percentage of a group finding the test acceptable), whereas qualitative studies can provide insights into why a particular test may be more or less acceptable than another and under what circumstances.

4.7.6. Feasibility

Key message 28 – Investigate the feasibility of implementing the test

The feasibility of a novel test or the ability of programmes to correctly implement it can have important implications for recommendations and the uptake of that test. Potential barriers to implementation across relevant settings from the perspective of patients, providers and policymakers should be measured directly through stakeholder surveys, qualitative research or other methods; ideally, evidence should be generated on how those barriers can be addressed. Elements that deserve consideration include maintenance and calibration requirements and support systems, reagent kit storage/stability, specimen preparation steps, resources required, cost to the programme, quality of training materials, trained human resource needed, connectivity, overall ease of use, infrastructure and space requirements, specimen transport requirements, biosafety considerations, availability of other assays to use on the same instrument (for multi-disease testing with assays for other conditions may be assessed by the relevant disease programmes within WHO), and an instrument's physical footprint.

[38] Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. *BMC Health Services Research.* 2017;17(1):88. doi:10.1186/s12913-017-2031-8.

Why this is important

The feasibility of diagnostic intervention refers to the likelihood that it can be implemented in health systems and properly carried out by the intended implementers (i.e., healthcare works or skills laboratorians) in a particular context; therefore, it is typically based on evidence and deliberations about enablers and barriers to implementation. Evidence for addressing gaps is important in the GDG process. This evidence is captured in review summaries and is often used to inform implementation considerations and further research sections within WHO guidelines as well as WHO operational handbooks.

4.8. Beyond initial WHO guideline development: Evidence to change or strengthen WHO recommendations

WHO policy decisions and recommendations for use of TB diagnostic interventions are evidence-based. Initial policy decisions on a new intervention often rely heavily on evidence from studies on diagnostic accuracy that are conducted in controlled (non-routine) settings. Based on this evidence and considering the broader evidence base on how TB diagnostics may affect decision-making, judgements on patient-important outcomes are made (see section 3 and section 4). At this stage, information on outcomes beyond accuracy (see section 4.5) as well as on cost, feasibility, acceptability, and possible effects on equity (see section 4.6) is often limited.

Box x. Where to find evidence gaps within WHO documents

Further evidence needs that are associated with each WHO recommendation on diagnostic testing may be referenced in the respective implementation considerations, monitoring and evaluation, and further research subsections in the WHO consolidated guidelines on tuberculosis – Module 3: Tuberculosis. Reviewing judgements of the certainty of evidence and reasons for downgrading can be directly informative on what additional evidence may be needed.

Initial recommendations in favour of a diagnostic intervention are therefore frequently conditional (not strong) as they are based on very low, low, or moderate (not high) certainty of the underlying evidence (see section 3.3). The types of evidence that are most frequently needed and may have the greatest impact to strengthen an initial conditional recommendation include:

- Further diagnostic accuracy evidence that helps to increase the certainty of the evidence altogether on diagnostic performance, in particular if this was a limiting factor of the initial recommendation (see section 3.3);
- direct evidence on patient-important outcomes (both intermediate and final), in particular if uncertainty about the indirectness of using diagnostic accuracy to support recommendations was a limiting factor of the initial recommendation (see sections 3.6 and 4.4);
- evidence on feasibility, acceptability, resources required, cost-effectiveness and effects on equity of the intervention in the intended setting of use (see section 4.5).

Further, initial recommendations are often limited in scope in terms of the indication of a new test (e.g. limited to certain populations or specimen types for which evidence is available). Studies that target these evidence gaps are commonly useful to strengthen or broaden WHO recommendations. The types of evidence that are most frequently needed and may have the greatest impact to broaden the scope of an initial recommendation, depend on the evidence gaps identified during the initial guideline development process, but include evidence of diagnostic accuracy for initially

excluded specimen types and populations of interest – particularly populations that may benefit most from receiving the intervention (i.e., non-sputum assay use for those unable to produce sputa for testing).

Active engagement of key stakeholders, particularly national TB programmes and affected communities but also implementers, laboratory networks, researchers, and industry partners, is essential at this stage. Their input helps to ensure that efforts to generate additional evidence is relevant to programmatic contexts, and the needs of those that new tools and strategies are meant to serve. Such engagement also facilitates country-level uptake and promotes alignment between global guidance and local health system capacities.

5. Other relevant WHO processes

5.1. WHO's prequalification

WHO's prequalification of IVDs is coordinated through WHO's Prequalification Unit. Focus is placed on IVDs for priority diseases and their suitability for use in resource-limited settings. WHO's prequalification of IVDs is a comprehensive quality assessment of individual IVDs through a standardized procedure aimed at determining whether a product meets WHO's prequalification requirements.

The prequalification assessment process includes the following components:

- review of a product dossier;
- manufacturing site(s) inspection;
- labelling review; and
- external performance evaluation.

Products submitted for WHO's prequalification assessment that meet, as determined by WHO, WHO's prequalification requirements are included in WHO's list of prequalified IVDs. The duration of the validity of the prequalification status of a product is dependent on the manufacturer's fulfilment, within the applicable deadlines, of its post-qualification obligations and requirements. The findings of WHO's prequalification assessment are used to evaluate the safety, quality and performance of IVDs for the purpose of providing guidance to interested United Nations (UN) agencies, relevant intergovernmental or international organizations, and WHO Member States in their procurement decisions.

5.2. Technical Advisory Group (TAG)

WHO established a Technical Advisory Group (TAG) on Tuberculosis Diagnostics and Laboratory Strengthening in 2021. The TAG is composed of up to 25 members with a range of technical knowledge, skills, and experience in clinical laboratory sciences, TB diagnostics and global, regional, or country-level laboratory systems strengthening, including experts from ministries of health, national TB programmes, public health, academic and research institutions, and other partners. Members are purposefully selected and appointed by WHO following an open call for experts. Appointees serve in their personal capacities without renumeration for 3-year terms that are eligible for reappointment[39]. As an advisory body to WHO, the group:

- advises WHO on priorities for TB diagnostic strategies that are identified by the WHO Secretariate in response to Member State needs and in line with the work of the wider WHO Strategic and Technical Advisory Group for Tuberculosis (STAG-TB); and
- provides rapid, independent evaluation and advice to WHO on scientific and technical aspects of TB diagnostic tools, technologies, methods and approaches which cannot be addressed within the scope of established WHO guideline development processes. This includes interim assessment of evidence on the use of new within-class diagnostic technologies for which a WHO prequalification process has not yet been established (see Section 1.3).

39 Terms of Reference for the Technical Advisory Group on Tuberculosis Diagnostics and Laboratory Strengthening

5.3. Expert Review Panel for Diagnostics

The Expert Review Panel for Diagnostics (ERPD) is a quality assurance mechanism used to assess the risks and benefits associated with the procurement and use of IVD medical devices that may have a substantial public health impact, but lack WHO recommendations, are not in the scope of prequalification or have not yet been prequalified or undergone stringent regulatory assessment by a founding member of the Global Harmonization Task Force. The ERPD approval is an interim, time-limited mechanism that aims to facilitate early access to innovative IVDs, provided the potential benefits significantly outweighs the risks associated with their use.

The ERPD is an independent advisory group of technical assessors to assess whether candidate IVDs meet specified safety, quality and performance expectations and determines their risk category in order to support procurement decisions.

The organization/programme requesting the ERPD review are the respective owners of the ERPD rounds and are responsible for launching the ERPD calls for expression of interest and communication with manufacturers.

To be eligible for ERPD review, an IVD must meet the state-of-the-art quality standards and performance criteria as defined in the call for expression of interest and be manufactured at a site that is compliant with ISO 13485: 2016 Medical devices – Quality management systems – Requirements for regulatory purposes.

5.4. WHO essential diagnostics lists

The WHO Essential Diagnostics List (EDL)[40] is an evidence-based register of IVDs that supports countries to facilitate their decision-making processes for selection and procurement of diagnostics. The EDL also provides a policy framework to support countries in their efforts to establish national lists that are tailored to individual settings. The WHO EDL publication is accompanied by an electronic EDL (eEDL[41]) that is an open access database of IVDs incorporating updates from each version and allowing users to search diagnostics by name, indication or test purpose and filter the complete list by disease/ health condition, setting, assay format, IVD purpose, specimen type, and year of WHO recommendation. Categories of WHO-recommended TB diagnostics are included in both the EDL and eEDL with indications for the lowest level of use within the health system, the intended target of test detection (i.e., DNA, DNA resistance mutations, mycobacteria), and test format (i.e., rapid diagnostic test, immunoassay, molecular line probe assay). The list is updated iteratively and the most recent version may be accessed on the relevant WHO website[42].

5.5. WHO Coordinated Scientific Advice procedure

The WHO Coordinated Scientific Advice (CSA) procedure is a single-entry service that lets developers of diagnostics (and other priority health products) obtain a joint, written assessment of their

[40] World Health Organization. The selection and use of essential in vitro diagnostics: report of the fourth meeting of the WHO Strategic Advisory Group of Experts on In Vitro Diagnostics, 2022 (including the fourth WHO model list of essential in vitro diagnostics). Geneva: World Health Organization; 2023. (WHO Technical Report Series No. 1053). ISBN: 978-92-4-008109-3

[41] World Health Organization. WHO Model List of Essential In Vitro Diagnostics (EDL) — Tuberculosis (TB) test categories. Geneva: World Health Organization

[42] World Health Organization. Selection, access and use of in vitro diagnostics. In: Health products policy and standards. Geneva: World Health Organization

development plans from both the relevant WHO technical department and the WHO PQ Team. By clarifying in advance the evidence WHO will later need for PQ listing and guideline formulation, CSA is intended to shorten the journey from late-stage research to large-scale public-health use.

Developers may request CSA once clinical development is under way but before the definitive clinical-validation study for IVDs is finalised. Provided the product shows significant public-health value and normally falls within PQ's remit, WHO conducts a brief eligibility screen, reviews a fuller submission, holds an online meeting to clarify outstanding issues, and issues a consolidated advice letter. When the dossier is complete, the four-step cycle—eligibility, dossier review, meeting and written report—usually lasts about ten weeks.

The advice covers clinical, analytical, quality-manufacturing and implementation considerations, aligning development plans with existing WHO Target Product Profiles. It remains non-binding, i.e. the CSA is not a pre-evaluation and does not predetermine the outcome of any future PQ application or guideline review. Nonetheless, engaging early through CSA can help developers design studies that will satisfy WHO's later evidence requirements, minimising costly re-work at the PQ or guideline stage.

References

Annexes

Please note that Annexes are still in early draft with internal review not fully completed.

Annex 1: Additional information relating to introductory material

Table A1 Existing WHO target product profiles for diagnostic testing across the TB clinical spectrum

Indication	WHO TPP document
Detection of TB infection	-
Detection of progression from TB infection to disease*	Target Product Profile (TPP) and a framework for evaluation for a test for predicting progression from tuberculosis infection to active disease
Screening for TB disease	Target product profiles for tuberculosis screening tests
Diagnosis of TB disease	Target product profile on a rapid test for detecting M. tuberculosis at the peripheral level
Diagnosis of TB drug resistance	Target product profile on next-generation DST for M. tuberculosis at the peripheral level
TB treatment monitoring	Target product profiles for tests for tuberculosis treatment monitoring and optimization

** Science and terminology has evolved significantly since this document was published and so this is no longer considered up to date.*

Table A2. Differences and similarities in scope and approach between WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections assessment of TB tests, WHO prequalification and regulatory approval

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
Prerequisite for	Identified public health need and new products developed that are	• Applications for WHO's prequalification assessment of an IVD are only accepted for	The manufacturer of an IVD is expected to design and manufacture a product that is safe and

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
evaluation	design-locked and available on the market.	products that are found by WHO to meet the below eligibility criteria: • The Expression of Interest for WHO's performance evaluation of that IVD has been received and accepted by WHO (where applicable); and • The product must be manufactured by the original product manufacturer (i.e., rebranded products are not accepted); and • Applications must be submitted by the original manufacturer of the product (i.e., applications from a rebrander are not accepted); and • The products must be in design lock-down when the application is submitted for WHO's prequalification assessment; and • The product must have been validated by the manufacturer and the established performance claims are included in the IFU. • In addition, WHO reserves the right to determine eligibility for WHO's prequalification assessment of an IVD considering the product categories for which there exist few other prequalified products[3].	performs according to established standards throughout its life-cycle. The harmonized Essential Principles3 should be fulfilled in the design and manufacturing of IVDs to ensure that they are safe and perform as intended. An application is required providing evidence demonstrating that applicable requirements for labelling, manufacturing quality systems, and performance are met and potential risks to patients or users are minimized. Specific requirements depend on the risk classification of the test. Some national regulatory authorities may require local clinical performance data. An application is also required to sell a device to a qualified investigator for the purpose of conducting investigational testing/clinical trials in human subjects. Specific requirements also depend on the risk classification of the test.
Goal	To provide guidance on use of a specific class of diagnostic technologies considering a systematic review of evidence on its impact on patient important outcomes, diagnostic accuracy, economic evidence, feasibility, accessibility, and equity in specified populations against an appropriate	The purpose and objective of WHO's prequalification of IVDs are to independently assess the safety, quality and performance of IVDs for the purpose of providing guidance to interested UN agencies, relevant intergovernmental or international organizations, and WHO Member States in their procurement decisions.	The main goal is to ensure that diagnostic tests are safe, perform according to established standards and meet quality standards before they are authorized for import or sale or advertisement for sale. The IMDRF Essential Principles of Safety and Performance of Medical Devices and IVD Medical

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
	comparator.		Devices provide a common set of fundamental design and manufacturing requirements for medical devices that, when met, provide assurance the device is safe and performs as intended, offers significant benefits to, among others, manufacturers, users, patients/consumers, and to Regulatory Authorities. In the case of investigational testing/clinical trials authorizations, the main goal is to ensure the investigational device can be used without seriously endangering the life or health of patients, users or other persons, that the testing is not contrary to the best interests of the patients and that the objective of the testing is achievable
Meaning of a decision	WHO recommendations used by Member States to inform the selection and use of a new diagnostic intervention (i.e., technology, sample, or strategy) for an intended purpose (i.e., detection of TB infection, disease, drug resistance) in a specified population. A WHO recommendation makes the test eligible for Global Fund grants and procurement via GDF, UN agencies, governments and other donors.	UN agencies, international or intergovernmental procurement organizations and/or WHO Member States may use WHO's list of prequalified IVDs to inform their respective procurement decisions.	A test is deemed licensed or approved for the purposes of its importation, sale or advertisement. A positive decision (licensed/approved test) means that the test is expected to perform as intended by the manufacturer and shall be effective for the medical conditions, purposes and uses for which it is manufactured, sold or represented. In the case of investigational testing authorizations, if a device is authorized, the manufacturer or importer may sell the device to a qualified investigator for the purpose of conducting investigational testing/clinical trials.
Remit	Global - UN agencies, international	Global - UN agencies, international or	National or Regional, depending on the regulatory

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
	or intergovernmental procurement organizations and/or WHO Member States may use WHO's recommendations to inform TB diagnostic selection, procurement, and use in countries.	intergovernmental procurement organizations and/or WHO Member States may use WHO's list of prequalified IVDs to inform their respective procurement decisions.	system. 4
Main criteria affecting decision-making	Criteria affecting decision-making by regulatory authorities may be considered but the main criteria are the so-called EtD criteria (Section x.x): • certainty of evidence • values • balance of desirable and undesirable effects • resources required and cost–effectiveness • equity • acceptability • feasibility	WHO's prequalification of IVDs independently assesses the safety, quality and performance of IVDs. The prequalification assessment process includes the following components: • review of a full product dossier; • manufacturing site(s) inspection; and • labelling review.	Decision-making based on information submitted meeting regulatory requirements of safety, effectiveness and quality, including labelling, quality management and manufacturing, analytical and clinical studies. In the case of test for investigational testing, the decision-making is based on information submitted meeting regulatory requirements for investigational use.
Mechanism to ensure reliability and quality of evidence	Early discussion with WHO technical departments (e.g. the Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections) is encouraged. Additionally, the CSA procedure is available (Section x.x). Systematic and transparent review of evidence based on the GRADE framework, including the use of	WHO will perform the prequalification assessment as per published procedures and requirements. The information submitted in the product dossier is reviewed and assessed by external experts (assessors) selected and appointed by WHO. Assessors involved in the product dossier review have appropriate qualifications and expertise in the relevant fields, are required to comply with the confidentiality and conflict of interest rules of WHO,	Publication of list of recognized standards, guidance documents and applicable notice to industry. Pre-submission meetings are encouraged. Review of evidence against recognized diagnostic, laboratory and/or technical guidance documents and peer-reviewed published literature.

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
	evidence synthesis, evidence appraisal and management of conflicts of interest.	and act as temporary advisers to WHO. Each manufacturing site(s) inspection is performed by an inspection team on behalf of WHO. WHO's inspection team is typically composed of a WHO staff inspector and external experts (also called "co-inspectors") selected and appointed by WHO. The external experts involved in the manufacturing site inspection are expected to have appropriate competence, qualifications and expertise in the relevant fields; and will be required to comply with the confidentiality and conflict of interest rules of WHO	
Evidence base for evaluation of benefits and harms	Systematic review of all available evidence relating to intervention impact on patient important outcomes, diagnostic accuracy, economic evidence, feasibility, accessibility, and equity.	The prequalification assessment process includes the following components: • review of a full product dossier; • manufacturing site(s) inspection; and • labelling review.	Review of the evidence on which the manufacturer relies to ensure that the device meets the applicable regulatory requirements. The scientific evaluation takes into consideration relevant information available (guidance documents (e.g. CLSI), clinical guidelines, international standards, etc.)[5]
Approach to analysis and decision-making	Systematic review of all available evidence collected through targeted and public calls for data, followed by meta-analysis, and assessment by a WHO guideline development group or technical advisory group whose membership considers gender balance, geographic representation, stakeholder representation, and relevant technical expertise.	WHO will take the prequalification assessment decision (whether positive or negative) regarding a product only after: 1. all components of WHO's prequalification assessment of the IVD (i.e., product dossier review, manufacturing site(s) inspection and labelling review) have been completed; and 2. if the IVD undergoing WHO's prequalification assessment is also required to undergo WHO's performance evaluation, WHO's performance	Conducting an analysis of safety and effectiveness or risk/benefit analysis based on the review of the available evidence. Typically, the National Regulatory Authority or designated body will assess the evidence submitted in the regulatory submission (e.g. a product dossier) and will assess the Quality Management System either through an on-site inspection or a desktop review of QMS documentation. [6] In the case of tests for investigational testing, an

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
		evaluation of the IVD has been completed.	analysis of safety and suitability for the proposed clinical study is conducted based on the review of the information provided (background, Information, risk assessment, Ethics Committee or IRB Approval(s), protocol, device label, investigator agreements).
Considerations after recommendation and approval	WHO recommendation process results in identification and issuance of implementation considerations, research priorities, and opportunities for operational research to support informed implementation and improve the strength of future recommendations. WHO's remit includes operational assistance and facilitation of implementation of recommended interventions. Guideline recommendations continually evolve based on reassessment of existing and novel tests. Minor version changes of tests are not usually subject to the guideline development process and are evaluated using other processes that are outside the scope of this	If the product is included in WHO's list of prequalified IVDs, the manufacturer will be responsible for timely and fully meeting its post-qualification obligations, namely: • prequalification commitments; and • annual reporting to WHO; and • reporting of changes to WHO; and • post-market surveillance obligations; and • undergoing manufacturing site(s) inspections; and • ongoing compliance with WHO's prequalification technical specifications where these exist; and • payment of the annual fee.	Typically, mature regulatory systems include provisions embracing the product's life cycle, such as post-market surveillance and vigilance systems. Compliance with any terms and conditions imposed at licensing/approval.

	WHO Department for HIV, Tuberculosis, Hepatitis and Sexually Transmitted Infections **Assessment**	**WHO Prequalification**	**International or National Regulatory approval***
	document.		

*Requirements for regulatory approval differ between countries and this table outlines what is mostly consistent across countries and organizations, including e.g. IMDRF guidance.

EtD : evidence to Decision framework

Annex 2: Additional information relating to methodology for development of GEG

TBD if anything needs to be added here

Annex 3: Additional information relating WHO guideline development process and the GRADE approach

A3.1 Comparative and non-comparative guideline questions

Guideline questions on new tests can be formulated either as non-comparative or as comparative questions. Comparative questions can help optimize diagnostic algorithms, assessing the new test/strategy considering the existing practise and weighing benefits and harm. Whereas non-comparative questions are more contextual and are considered when there is no comparator or evidence available on head-to-head comparisons. Table x summarizes key differences between these guideline questions.

Table A3.1. Advantages and use of comparative and non-comparative evidence *[NOTE: This is an early draft with internal discussions not concluded but input welcome]*

	Non-comparative	Comparative
Advantages	• Possible even if no comparator exists or if no evidence on comparators is available	• Responds directly to the question of whether or in what context the diagnostic intervention should replace or complement the current standard of care
Most appropriate when	• No comparator available (i.e. new indication) • No data on direct head-to-head comparison between diagnostic intervention and comparator available • The need exist to support market competition • The need exist to preserve earlier recommendations	• A clear comparator (or small number of comparators) exists that is relevant for most settings • Data on direct head-to-head comparisons is available for the majority of data points

A3.2 Diagnostic accuracy combined with other evidence

Diagnostic accuracy is an important component for any diagnostic intervention and has been described in detail in section 4.2. Beyond accuracy, an assessment of desirable and undesirable effects is also often based on a combination of test accuracy, together with other data (also called the diagnostic-accuracy-based approach). If test accuracy is used during guideline development, best estimates of the diagnostic accuracy are used in combination with prevalence estimates to compute the number of true-positives (TPs), true-negatives (TNs) –representing desirable effects—and false-positives (FPs) and false-negatives (FNs) –representing undesirable effects.

These estimates should be contextualized with evidence on the test's direct effects (such as procedural risks), TB natural history, the effectiveness of available treatment options, and the extent to which test results reliably guide clinical management. This comprehensive approach enables a more meaningful assessment of a diagnostic intervention's real-world impact. Further

methodological guidance is provided in Section 4.3 (diagnostic accuracy) and Section 4.4 (linkage between test results and management decisions).

Box x. ***Factors considered when the assessment of desirable and undesirable effects is based on a combination of test accuracy, together with other data (also called the diagnostic-accuracy- based approach)***

Accuracy
See section 4.3 for detailed guidance

Prevalence
Typically, a range of prevalence rates, based on evidence of TB prevalence in setting of intended use, is considered.

TPs, TNs, FPs and FNs
Computed based on accuracy estimates and prevalence estimates.

Evidence of test's effects
This criterion within the EtDs pertains to any *direct* effects of taking the specimen or doing the test on the patient (i.e. procedural harms) and thus there are typically, few/none expected (beyond from sampling). Typically, we have no included studies. We suggest capturing some information on this as part of studies, especially if the process differs from that used to obtain specimens for tests used for standard of care.

Evidence of management's effects
This criterion within the EtDs pertains to the effectiveness of available treatment options and the extent to which this improves over the natural history of TB. Relevant guidelines and literature can be referred to in the EtDs for this criterion. Some consideration should be given to differences in patient spectrum if new tests pick up patients where natural history or treatment effectiveness may differ from those that were in available studies.

Evidence on linkage between test results and management
See section 4.4

A3.3 Assessing the certainty of the evidence

Fig xx provides an overview of the guideline development process.

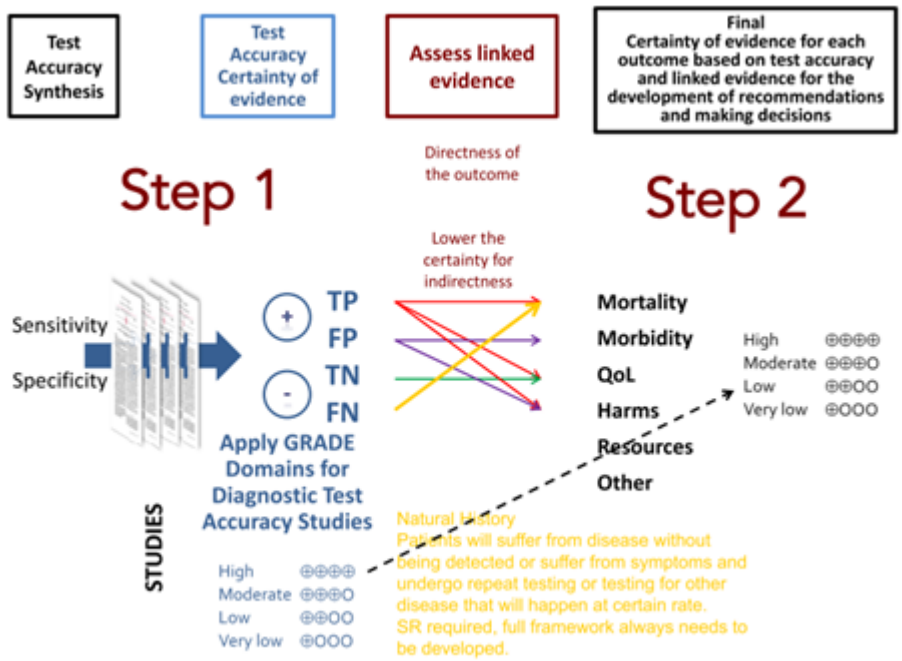


Fig. 4. Linking test accuracy to patient-important outcomes.

Source:

The five domains that could potentially downgrade the certainty of evidence for test accuracy are:

Risk of bias

Base your GRADE judgment on the QUADAS (QUADAS-2 or QUADAS-C) assessment. In general, if most judgments are low risk of bias in all four QUADAS-domains, judge risk of bias as not serious. Threshold to define what percentage will be acceptable to not downgrade for risk of bias, or downgrade one level or two levels should be decided *a priori*.

Indirectness

Directness of evidence will be how closely the included population, diagnostic intervention, outcome measures are to the research question. This is synonymous with applicability and generalizability.

For both sensitivity and specificity, note important differences between the populations studied (prior testing, the spectrum of disease, and comorbidities), the setting, diagnostic intervention, and reference standards, and assess whether differences are sufficient to lower the certainty of evidence.

Indirectness could occur if the setting in which the test was done is not the intended use setting. Prevalence may be a rough gauge (and a surrogate for spectrum of disease) about whether there is indirectness in the populations studied. It is important to assess, 'Is the average or median prevalence in the included studies similar to the level found in practice, i.e. within the range of the three prevalence values provided in the GRADE evidence profile or summary of findings table.

Inconsistency

Inconsistency can be caused by clinical or methodological heterogeneity, or it may be unexplained. As GRADE recommends downgrading for unexplained inconsistency in sensitivity and specificity estimates, systematic review authors should state if they carried out pre-specified analyses, e.g. subgroup analyses or meta-regression, to investigate potential sources of heterogeneity and consider downgrading when they cannot explain inconsistency in the accuracy estimates.

Ideally, inconsistency should be assessed by using clearly defined thresholds that either resemble healthcare practice or will be used to guide practice. One way to visually assess inconsistency is by looking at forest plots, or 95% prediction regions on summary ROC plots (if available)

Imprecision

We consider a precise estimate to be one that would allow a clinically meaningful decision. It is important to consider width of CIs and sample size.

Additionally, it would be important to use the actual TP, FP, FN, TN for the three prevalence values to assess if the clinical decision will change based on these numbers for these prevalence values. Additionally, based on the CI approach, GRADE recommends that, prior to rating, systematic review authors consider defining judgment thresholds for a very accurate, accurate, inaccurate, and very inaccurate test. When a CI appreciably crosses the predefined judgment threshold(s), one should consider rating down certainty of evidence by one or more levels, depending on the number of thresholds crossed. When the CI does not cross judgment threshold(s), GRADE suggests considering the sample size for an adequately powered test accuracy review. It is important to note that inconsistency and imprecision are related so avoid double counting and downgrading twice.

Dissemination bias

Selective publication of studies based on the nature or direction of their findings, often favoring those with higher accuracy estimates, could lead to publication bias and can distort overall assessment of a test's performance in meta-analyses. This can lead to downgrading of evidence. However, unlike intervention studies, DTA studies are particularly prone to threshold effects and variability in study design, making the detection of publication bias more complex. In diagnostic studies, statistical tests for funnel plot asymmetry (e.g., Begg, Egger, Harbord, Peters) are not appropriate, as they can falsely suggest publication bias when odds ratios are large. Deek's test may be used but suffers from low power when there is heterogeneity. Therefore, any asymmetry should be interpreted cautiously, considering alternative explanations such as study size or patient characteristics rather than assuming publication bias[43].

GRADE CERQual

Sections 3.4.1–3.4.5 and most of this document in general are focused on quantitative evidence, to which GRADE applies. Qualitative research evidence can add value or complement quantitative evidence, especially where there is a more in-depth understanding on the question of *why* things are the way they are, rather than *how much* they are a certain way (e.g. why something is acceptable or feasible rather than to what degree people find something acceptable or feasible). GRADE CERQual is a transparent and structured approach for assessing how much confidence to place in individual

[43] Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.

review findings (i.e. to assess the extent to which the review finding is a reasonable representation of the phenomenon of interest) (15). The review findings are the results of a qualitative evidence synthesis and can be presented in different formats (e.g. a theme, category, thematic framework, theory or contribution to theory) and at different levels (e.g. descriptive or aggregative and interpretive or narrow; for example, in relation to a specific health care setting or more broadly cutting across several different kinds of social care settings). At least two members of the review team will arrive at CERQual assessments for each review finding through discussion of four key components, with equal weight given to each component:

- methodological limitations of included studies;
- coherence of a review finding;
- adequacy of data; and
- relevance of included studies to the review question.

Overall certainty of evidence

The GRADE approach requires the guideline panel to rate certainty of evidence separately for each outcome based on the evidence available and section 3.4.2 provides details on the five domains which might decrease the certainty of evidence. It is important to note that there are factors that can increase the certainty of evidence as well. These factors are large magnitude of effect, effect of plausible residual confounding, dose-response gradient. However, this phenomenon of increasing the certainty of evidence has not been observed in diagnostic evidence reviews.

To determine the overall certainty in effect estimates, it is important to consider only those outcomes that have been deemed important and critical by the guideline panel. If the certainty of evidence is same across all relevant outcomes, then that becomes the overall certainty of evidence. However, if the certainty of evidence differs across critical outcomes, the overall certainty is guided by the lowest certainty of evidence for any important and critical outcome.

It is important to note that in some instances, an outcome deemed critical at the beginning may change based on the evidence received, which could also change the overall certainty and recommendations. These are judgement calls made by the panel and are probably rare.

Certainty of the evidence of test's effects, management's effects and test result/management

The certainty of evidence concerning test effects, management effects and test results and management is appraised using the same criteria as for health outcomes (Section xx). It is important to note that evidence on all these indicators could vary significantly based on the settings. Also, the certainty of evidence may differ across different populations and country contexts.

Certainty of the evidence of test's effects

What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?

Certainty of the evidence of management's effects

What is the overall certainty of the evidence of effects of the management that is guided by the test results?

Certainty of the evidence of test result/management

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

Certainty of evidence of required resources

The certainty of evidence concerning required resources is appraised using the same criteria as for health outcomes (Section xx). It is important to note that evidence on resources could vary significantly based on the settings and resources considered. Also, the quality of evidence may differ across different resources. Evidence of actual resource use is generally preferable to indirect estimates of the costs of those resources (16). Pooling resource estimates from different studies is seldom done and should be carefully considered. However, pooling could be considered if there the outcome measures across studies have the same meaning and the estimates have been adjusted for geographical and temporal differences.

A3.4 Preparing evidence profiles and summary of findings tables

Evidence profiles are tables that display the ratings of the certainty of evidence together with summary effect estimates in a standardized format; summary of findings are tables that show abbreviated versions of the evidence profiles. These tables are a core element of the guideline development process. They represent the main format in which evidence is presented to the GDG members, to support their judgements about the magnitude of desirable and undesirable effects.

Figure A3.2 Example of evidence profile when using the diagnostic-accuracy-based approach

Question: Should LC-aNAAT on respiratory samples be used to diagnose PTB in children with signs and symptoms or screened positive for PTB, CRS?

Sensitivity	0.23 (95% CI: 0.20 to 0.27)
Specificity	1.00 (95% CI: 1.00 to 1.00)

Prevalences	1%	5%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of5%	pre-test probability of10%	
True positives (patients with PTB)	10 studies 1626 patients	cross-sectional (cohort type accuracy study)	serious[a]	not serious	not serious	not serious	none	2 (2 to 3)	12 (10 to 14)	24 (20 to 27)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having PTB)								8 (7 to 8)	38 (36 to 40)	76 (73 to 80)	
True negatives (patients without PTB)	10 studies 1197 patients	cross-sectional (cohort type accuracy study)	serious[a]	not serious	not serious	not serious	none	990 (987 to 990)	950 (947 to 950)	900 (897 to 900)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having PTB)								0 (0 to 3)	0 (0 to 3)	0 (0 to 3)	

Explanations

a. In 50% of the studies included in the meta-analysis, the composite reference standard was not considered independently of the index test.

Figure A3.3 Example of evidence profile when using the patient-important-outcomes-based approach

Table 1.: Xpert MTB/RIF compared to smear microscopy in adults with signs and symptoms of pulmonary tuberculosis

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Xpert MTB/RIF	smear microscopy	Relative (95% CI)	Absolute (95% CI)		
Mortality												
5 [1,2,3,4,5]	randomised trials	not serious [a]	not serious [b]	not serious	serious [c]	none	248/5265 (4.7%)	292/5144 (5.7%)	**RR 0.88** (0.73 to 1.05)	**7 fewer per 1,000** (from 15 fewer to 3 more)	⨁⨁⨁◯ MODERATE	CRITICAL
Cure												
2 [3,6,7]	randomised trials	not serious	not serious	not serious [d]	not serious	none	1786/2500 (71.4%)	1443/2080 (69.4%)	**OR 1.09** (1.02 to 1.16)	**18 more per 1,000** (from 4 more to 31 more)	⨁⨁⨁⨁ HIGH	CRITICAL
Pre-treatment loss to follow up												
3 [3,4,5]	randomised trials	not serious	serious [3,4,5,e]	not serious	not serious	none	81/642 (12.6%)	95/523 (18.2%)	**RR 0.59** (0.42 to 0.84)	**74 fewer per 1,000** (from 105 fewer to 29 fewer)	⨁⨁⨁◯ MODERATE	IMPORTANT
Time to diagnosis												
2 [4,5]	randomised trials	not serious [a]	not serious	not serious [f]	not serious [g]	none	956 participants	968 participants	**HR 1.05** (0.93 to 1.19) [Time to diagnosis]	**5 more per 1,000** (from 7 fewer to 18 more)	⨁⨁⨁⨁ HIGH	CRITICAL
							-	10.0%		**5 more per 1,000** (from 7 fewer to 18 more)		

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Xpert MTB/RIF	smear microscopy	Relative (95% CI)	Absolute (95% CI)		
Time to treatment												
4 [1,3,4,5]	randomised trials	not serious [a]	not serious	not serious [f]	serious [h]	none	4055 participants	4153 participants	**HR 1.00** (0.75 to 1.32) [Time to treatment]	**0 fewer per 1,000** (from 24 fewer to 30 more)	⨁⨁⨁◯ MODERATE	CRITICAL
							-	10.0%		**0 fewer per 1,000** (from 24 fewer to 30 more)		
Mortality in HIV-positive participants												
2	randomised trials	not serious	not serious	not serious	serious [i]	none	66/1211 (5.5%)	75/1055 (7.1%)	**RR 0.76** (0.59 to 1.00)	**17 fewer per 1,000** (from 29 fewer to 0 fewer)	⨁⨁⨁◯ MODERATE	CRITICAL
New outcome												
									not estimable		-	

CI: Confidence interval; **RR**: Risk ratio; **OR**: Odds ratio; **HR**: Hazard Ratio

Explanations

a. For all randomized trials, blinding of physicians to what test was done was impossible since knowing which test was done is part of the intervention itself. For example, the Xpert test has higher sensitivity than smear microscopy (and also produces RIF resistance results) and physicians must be allowed to take this into account when deciding about patient management. While outcomes between patients may therefore be different due to lack of blinding this was not judged to be a source of bias but rather the mechanism through which the intervention had an effect. Outcome measurement could theoretically have been influenced by the lack of blinding but this was deemed unlikely to cause bias of important magnitude. Overall, the lack of blinding was therefore judged not to put studies at increased risk of bias.Type a message

b. No evidence of inconsistency, four studies in the direction of showing benefit.

c. The 95% CI is wide likely suggesting imprecision. We caution about interpreting non-significance as no effect when the CI likely includes an effect that may be clinically important. We downgraded one level for Imprecision.

d. Cure is the outcome of interest for patient important outcome. Studies have reported treatment success which includes those cured and those completing treatment without evidence for treatment failure . However, we did not downgrade for indirectness

e. Variability in time for assessment of pre-treatment loss to follow up; Churchyard 2015 assessed within 28 days after enrolment, Cox 2014 assessed by three months after enrolment and Theron 2014 assessed by the end of the study (six months)

f. The results are from trials that directly compared the populations, interventions and outcomes of interest. We did not downgrade for imprecision

g. The results suggest that Xpert did not improve time to diagnosis compared to smear microscopy but the direction of effect is towards benefit. We did not downgrade for imprecision because the 95% CI is narrow.

h. The results suggest that Xpert did not improve the time to treatment comapred to smear microscopy. The 95% CI is wide likely suggesting imprecision

i. Similarly, the 95% CI is wide likely suggesting imprecision. We caution about interpreting non-significance as no effect when the CI likely includes an effect that may be clinically important. We downgraded one level for Imprecision.

A3.5 Evidence to Recommendations

Once the evidence has been retrieved, it is summarized and rated for certainty after which the WHO convenes a meeting of the GDG, where summary of findings tables and other information are presented and discussed using a format of structured deliberation under the guidance of a guideline methodologist. The outputs of the discussions are captured in evidence to decision (EtD) tables, which show how the factors that determine the direction and strength of a recommendation inform the process of developing the recommendation. These tables enhance the transparency of the process, focus the discussions of the GDG and permit recording of the judgements made about each factor and how each one contributed to the recommendation. Table x.x explains the 17 EtD criteria typically evaluated as part of the overall assessment of the evidence.

In the following sections we will provide brief guidance on how evidence is reviewed, rating certainty and making judgements for these EtD criteria.

Table A3.4. Overview of the 17 EtD criteria typically evaluated as part of the overall assessment of the evidence[a]

EtD criterion (GEG section)		Signalling questions	Categories of judgements
1	**Problem**	Is the problem a priority?	o No o Probably no o Probably yes o Yes o Varies o Don't know
2 (4.3)	**Test accuracy**	How accurate is the test?	o Very inaccurate o Inaccurate o Accurate o Very accurate o Varies o Don't know
3 (4.2-4.4)	**Desirable effects**	How substantial are the desirable effects?	o Trivial to no effect o Small o Moderate o Large o Varies o Don't know
4 (4.2-4.4)	**Undesirable effects**	How substantial are the undesirable effects?	o Trivial to no effect o Small o Moderate o Large o Varies o Don't know
5 (3.7.2)	**Certainty of the evidence of test accuracy**	What is the overall certainty of the evidence of test accuracy?	o Very low o Low o Moderate o High
6 (3.8.2)	**Certainty of the evidence of test's effects**	What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?	o Very low o Low o Moderate o High
7 (3.8.2)	**Certainty of the evidence of management's effects**	What is the overall certainty of the evidence of effects of the management that is guided by the test results?	o Very low o Low o Moderate o High
8 (3.8.2)	**Certainty of the**	How certain is the link between	o Very low

	evidence of test result/management	test results and management decisions?	o Low o Moderate o High
9 (3.7.4)	**Certainty of effects**	What is the overall certainty of the evidence of effects of the test?	o Very low o Low o Moderate o High
10 (3.7.5 and 4.5.1)	**Values**	Is there important uncertainty about or variability in how much people value the main outcomes?	o Important uncertainty or variability o Possibly important uncertainty or variability o Probably no important uncertainty or variability o No important uncertainty or variability
11 (3.8.3)	**Balance of effects**	Does the balance between desirable and undesirable effects favour the intervention or the comparison?	o Favours the comparator o Probably favours the comparator o Does not favour either the intervention or the comparator o Probably favours the intervention o Favours the intervention
12 (4.5.2)	**Resources required**	How large are the resource requirements (costs)?	o Large costs o Moderate costs o Negligible costs and savings o Moderate savings o Large savings o Varies o Don't know
13 (3.8.4)	**Certainty of evidence of required resources**	What is the certainty of the evidence of resource requirements (costs)?	o Very low o Low o Moderate o High o No included studies
14 (4.5.3)	**Cost–effectiveness**	Does the cost–effectiveness of the intervention favour the intervention or the comparison?	o Favours the comparator o Probably favours the comparator o Does not favour either the intervention or the comparator o Probably favours the intervention o Favours the intervention o Varies o No included studies
15 (4.5.4)	**Equity**	What would be the impact on health equity?	o Reduced o Probably reduced o Probably no impact o Probably increased o Increased o Varies o Don't know
16 (4.5.5)	**Acceptability**	Is the intervention acceptable to key stakeholders, in relation to the comparator?	o No o Probably no o Probably yes o Yes o Varies

17 (4.5.6)	**Feasibility**	Is the intervention feasible to implement, in relation to the comparator?	○ Don't know ○ No ○ Probably no ○ Probably yes ○ Yes ○ Varies ○ Don't know

EtD: evidence to decision; GEG: guidance on evidence generation.

[a] A more detailed version of the table is provided in Annex x (Table Ax.x).

A3.6 Panel judgements

Judgement of the magnitude of desirable and undesirable effects

During evidence assessment, outcomes are referred to as desirable and undesirable based not on their inherent nature (e.g. death is undesirable, cure is desirable) but depending on whether the observed effects for a certain outcome favour the intervention or the comparator. Thus, outcomes for which effects favour the intervention will be listed as "desirable effects", whereas those that favour the comparator will be listed as "undesirable effects" within the EtD tables. The GRADE EtD framework then classifies effect sizes as trivial, small, moderate or large. This determination is made based on a collective judgement by the GDG; in some cases, this requires considerable deliberation. Judgements on the magnitude of desirable and undesirable effects are influenced by how guideline panels rate the effect sizes and the relative importance of prioritized outcomes.

For example, when test accuracy is used during guideline development, best estimates of the diagnostic accuracy are used in combination with prevalence estimates to compute the number of true-positives (TPs), true-negatives (TNs) –representing desirable effects—and false-positives (FPs) and false-negatives (FNs) –representing undesirable effects (see Annex 3 A3.4). These estimates should be contextualized with evidence on the test's direct effects (such as procedural risks), TB natural history, the effectiveness of available treatment options, proportion of inactionable results, and the extent to which test results reliably guide clinical management. Evidence on other relevant outcomes (e.g. time to result or effects on pre-treatment loss to follow-up) are considered as part of the desirable and undesirable effects, alongside evidence on diagnostic accuracy. This comprehensive approach enables a more meaningful assessment of a diagnostic intervention's real-world impact and in assessing the desirable and undesirable effects.

Judgement of the balance of effects

The balance of effects reflects the risk–benefit ratio of an intervention, considering the overall certainty of the evidence and how the outcomes are valued by those receiving it. It is thus based on the combination of judgements on the previous four EtD criteria (desirable effects, undesirable effects, certainty of effects and values). This judgement about the balance of effects is a strong determinant of the direction and strength of the final recommendation, even after considering the other important GRADE criteria.

Figure A3.5 Example of summary judgements across 17 EtD criteria when using the diagnostic-accuracy-based approach

	JUDGEMENT						
PROBLEM	No	Probably no	Probably yes	**Yes**		Varies	Don't know
TEST ACCURACY	Very inaccurate	Inaccurate	Accurate	**Very accurate**		Varies	Don't know
DESIRABLE EFFECTS	Trivial	Small	Moderate	**Large**		Varies	Don't know
UNDESIRABLE EFFECTS	**Trivial**	Small	Moderate	Large		Varies	Don't know
CERTAINTY OF THE EVIDENCE OF TEST ACCURACY	Very low	Low	Moderate	**High**			No included studies
CERTAINTY OF THE EVIDENCE OF TEST'S EFFECTS	Very low	Low	Moderate	High			**No included studies**
CERTAINTY OF THE EVIDENCE OF MANAGEMENT'S EFFECTS	Very low	Low	Moderate	**High**			No included studies
CERTAINTY OF THE EVIDENCE OF TEST RESULT/MANAGEMENT	Very low	Low	Moderate	**High**			No included studies
CERTAINTY OF EFFECTS	Very low	Low	Moderate	**High**			No included studies
VALUES	Important uncertainty or variability	Possibly important uncertainty or variability	Probably no important uncertainty or variability	**No important uncertainty or variability**			
BALANCE OF EFFECTS	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or the comparison	Probably favors the intervention	**Favors the intervention**	Varies	Don't know
RESOURCES REQUIRED	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	**Varies**	Don't know
CERTAINTY OF EVIDENCE OF REQUIRED RESOURCES	Very low	Low	Moderate	High			**No included studies**
COST EFFECTIVENESS	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or the comparison	Probably favors the intervention	Favors the intervention	Varies	**No included studies**
EQUITY	Reduced	Probably reduced	Probably no impact	Probably increased	**Increased**	Varies	Don't know
ACCEPTABILITY	No	Probably no	Probably yes	**Yes**		Varies	Don't know
FEASIBILITY	No	Probably no	Probably yes	**Yes**		Varies	Don't know

TYPE OF RECOMMENDATION

Strong recommendation against the intervention	Conditional recommendation against the intervention	Conditional recommendation for either the intervention or the comparison	Conditional recommendation for the intervention	**Strong recommendation for the intervention**

Figure A3.6 Example of summary judgements across 12 EtD criteria when using the patient-important-outcomes-based approach

	Judgement						
Problem	No	Probably no	Probably yes	**Yes**		Varies	Don't know
Desirable effects	Trivial	Small	**Moderate**	Large		Varies	Don't know
Undesirable effects	Large	Moderate	Small	**Trivial**		Varies	Don't know
Certainty of evidence	Very low	Low	**Moderate**	High			No included studies
Values	Important uncertainty or variability	Possibly important uncertainty or variability	Probably no important uncertainty or variability	**No important uncertainty or variability**			
Balance of effects	Favours the comparison	Probably favours the comparison	Does not favour either the intervention or the comparison	Probably favours the intervention	**Favours the intervention**	Varies	Don't know
Resources required	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	**Varies**	Don't know
Certainty of evidence of required resources	Very low	Low	Moderate	High			**No included studies**
Cost	Favours the comparison	Probably favours the comparison	Does not favour either the intervention or the comparison	**Probably favours the intervention**	Favours the intervention	Varies	No included studies
Equity	Reduced	Probably reduced	Probably no impact	**Probably increased**	Increased	Varies	Don't know
Acceptability	No	Probably no	Probably yes	**Yes**		Varies	Don't know
Feasibility	No	Probably no	**Probably yes**	Yes		Varies	Don't know

Type of recommendation

Strong recommendation against the intervention	Conditional recommendation against the intervention	Conditional recommendation for either the intervention or the comparison	Conditional recommendation for the intervention	Strong recommendation for the intervention
○	○	○	○	●

A3.7 Developing recommendations

Strong recommendations

When we can be very certain about the balance of effects (i.e. the desirable consequences clearly outweigh the undesirable consequences or vice versa, and the certainty is high or at least moderate), and other EtD criteria support this, WHO may issue a strong recommendation in favor of

or against an intervention. The implications of strong recommendations are that the recommendation can be adopted as policy directly by most Member States, most clinicians would follow it, most patients would want the recommended course of action and additional research is unlikely to alter the recommendation (17). A few paradigmatic situations where strong recommendations may be made despite the evidence being of low or very low certainty are outlined in Annex 1 (Table Ax.x).

Conditional recommendations

When we are uncertain about the balance of effects or where the balance may depend on circumstances specific to an individual or context (e.g. based on judgements on other EtD criteria), WHO will typically issue a conditional recommendation. The implication of conditional recommendations are that substantial debate may be required before the policy is adopted by Member States; clinicians will need to discuss different management options with each patient; most patients may want the recommended course of action, although some or even many may not; and additional research would be likely to strengthen and possibly alter the recommendation (17).

Conditional recommendation for either the intervention or the comparison

Guideline users benefit from clear recommendations. A conditional recommendation for either the intervention or the comparison should be reserved for rare situations when two alternative intervention options appear to have equivalent net desirable consequences across the EtD criteria after careful evaluation. This option should not be chosen if an intervention is compared with current practice or no intervention – this will not provide guidance and will often be meaningless. Furthermore, a conditional recommendation for either the intervention or the comparison may be based on a comparator that has a strong recommendation as a basis (e.g. if it was previously compared with no intervention); logically, this suggests that the new intervention would also be strongly recommended if compared with no intervention.

A3.8 Extrapolation

Extrapolation of evidence could be done when direct evidence from the target population or setting is lacking, but there is reasonable biological, clinical, or methodological justification to assume applicability. It should be considered when studies involve different populations (e.g., adults vs. children), healthcare settings, or test versions, provided key factors such as disease prevalence, test use, and care pathways are comparable. When extrapolating, GDG panel members should assess indirectness using tools like GRADE, evaluate the similarity of population, setting, test characteristics, and reference standards, and consider potential biases. Extrapolation should be guided by indirect evidence models where direct outcome data are missing, and assumptions must be clearly justified and transparently reported. Stakeholder input and context-specific feasibility, including resource availability and cultural relevance, should also be considered. See also section 4.2.5 on generating supporting evidence for extrapolation.

Annex 4: Current best practice and options & case studies

A4.1 Typical setup of tables to perform analyses

Diagnostic test accuracy study: one 2x2 table

		Reference standard	
		Pos	**Neg**
Index test	**Pos**	TP	FP
	Neg	FN	TN

➔ Based on this, the **sensitivity and specificity** of the index test can be estimated and reported together with 95%CIs.

Comparative diagnostic test accuracy study: two 2x2 table

Table 1: Among reference standard-positive individuals

		Comparator	
		Pos	**Neg**
Index test	**Pos**	Both TP	Test 1 FN Test 2 TP
	Neg	Test 1 TP Test 2 FN	Both FN

➔ Based on this, the **difference in sensitivity** can be estimated and reported together with 95%CI around the difference

Table 2: Among reference standard-negative individuals

		Comparator	
		Pos	**Neg**
Index test	**Pos**	Both FP	Test 1 TN Test 2 FP
	Neg	Test 1 FP Test 2 TN	Both TN

➔ Based on this, the **difference in specificity** can be estimated and reported together with 95%CI around the difference

Importantly, this way we are accounting for the paired nature of the data. The precision of the estimate influenced by the number of reference standard-positive/negative individuals (for difference in sensitivity/specificity, respectively) and the level of correlation between index test 1 and index test 2. Of course, sensitivity and specificity of both tests with 95%CIs can also be computed and should be reported as well.

Concordance study: one 2x2 table

		Index test 1	
		Pos	**Neg**
Index test 2	**Pos**	Concordant positive	Discordant
	Neg	Discordant	Concordant negative

➔ Based on this, the **% concordant-positive and %concordant-negative or overall concordance rates** can be estimated and reported together with 95%CI

Importantly, in the absence of a reference standard, it may be impossible to know how to interpret discordant results and for concordant results it is possible that both tests are giving consistently wrong results. Further, prevalence and we do not know the prevalence so particularly "overall concordance" could be strongly influenced by the underlying (unknown) prevalence and is not a useful measure.

A4.2 Reference standard

[NOTE: Putting this figure from one of the 2019 papers here. We may consider this or some adaptation or improved version for the annex alongside with a discussion outlining the pros and cons, based on what we already had in 2019.]

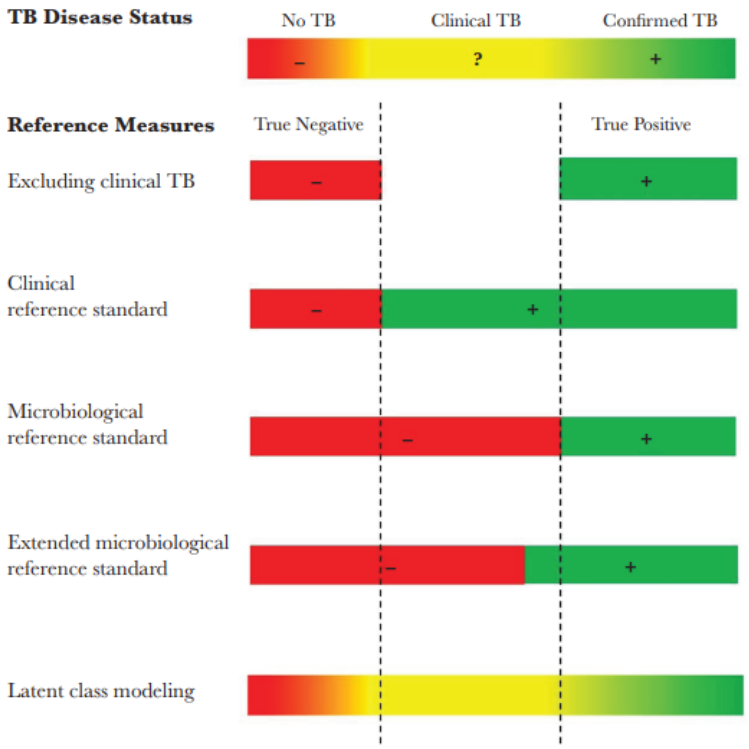


Figure 2. Defining outcomes for the reference standard test when designing tuberculosis diagnostic accuracy studies. For excluding clinical tuberculosis (TB), those with clinical TB (in yellow) are excluded from the analysis. For the clinical reference standard (CRS), those with clinical TB are often included as a confirmed TB case. For the microbiological reference standard (MRS), those with clinical TB are included as a noncase (no TB). The extended MRS classifies a few additional clinical TB cases into the confirmed TB category. Latent class modeling uses statistical modeling to incorporate all available test results to estimate the probability of TB in each individual, and test accuracy is estimated while accounting for uncertainty in disease classification [15, 16]. Ideally, results from all of the above should be presented but at least the "case/noncase," CRS, and MRS are easily done without much additional work.

[NOTE: Also from 2019, related to the above. Not suggesting we put it in main text but could consider this or something like it for annex]

Table 2. Sample Applications of Tuberculosis Reference Standards

Parameter	Excluding Clinical TB	MRS	eMRS	CRS
Tests and follow-up considered				
1–2 MGIT[a]	X	X	X	X
1–2 LJ[a]	X	X	X	X
Blood culture[a]	X	X	X	X
Urine Ultra	X	X	X	X
Sputum Xpert/Ultra	X	X	X	X
Additional testing[b]	…	…	X	X
Clinical follow-up	…	…	…	X
Symptoms and treatment at 2–3 mo				
Persistent symptoms	…	…	…	X
Initiation of ATT	…	…	…	X
Reference standard positive	Any of the tests considered is positive	Any of the tests considered is positive	Any of the tests considered is positive	Any of the tests considered is positive and/or TB treatment was started
Reference standard negative	None of the tests considered is positive and at least 1 test is negative	None of the tests considered is positive and at least 1 test is negative	None of the tests considered is positive and at least 1 test is negative	None of the tests considered is positive and the patient has no symptoms and TB treatment was not started
Unclassifiable (excluded from analysis)	Neither reference standard positive nor reference standard negative; patients with clinical TB	Neither reference standard positive nor reference standard negative	Neither reference standard positive nor reference standard negative	Neither reference standard positive nor reference standard negative

Abbreviations: ATT, anti-tuberculosis therapy; CRS, clinical reference standard; eMRS, extended microbiological reference standard; LJ, Löwenstein–Jensen; MGIT, mycobacterial growth indicator tube; MRS, microbiological reference standard; TB, tuberculosis.

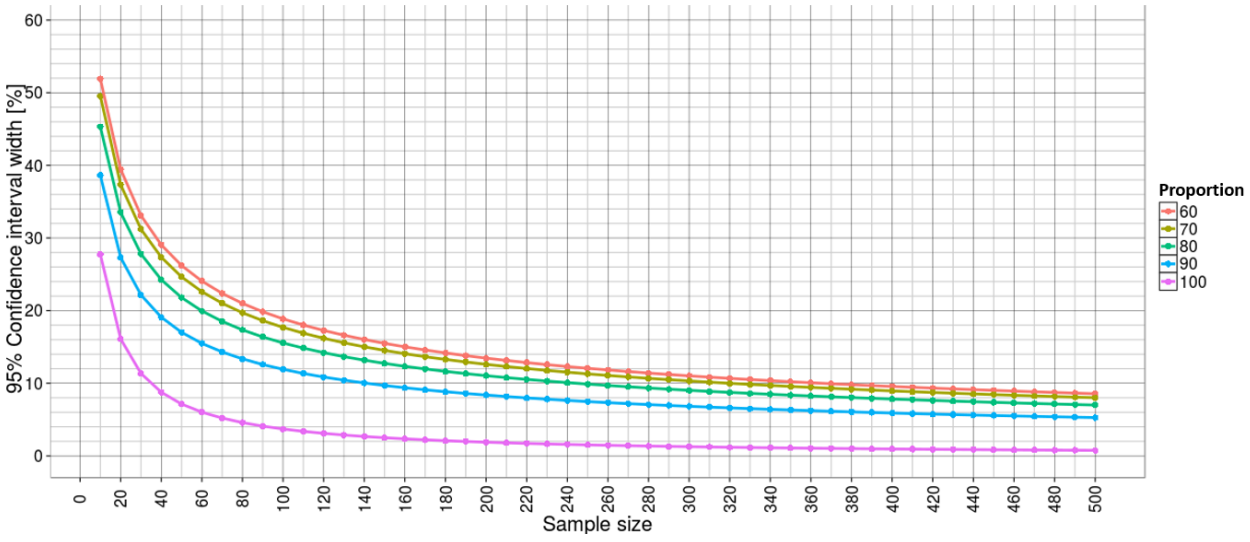
[a] *Mycobacterium tuberculosis* complex needs to be confirmed.

[b] Any additional mycobacterial culture or GeneXpert/Ultra from other respiratory and/or nonrespiratory samples (eg, pleural fluid, tissue biopsy, cerebrospinal fluid) that were performed based on clinical indication.

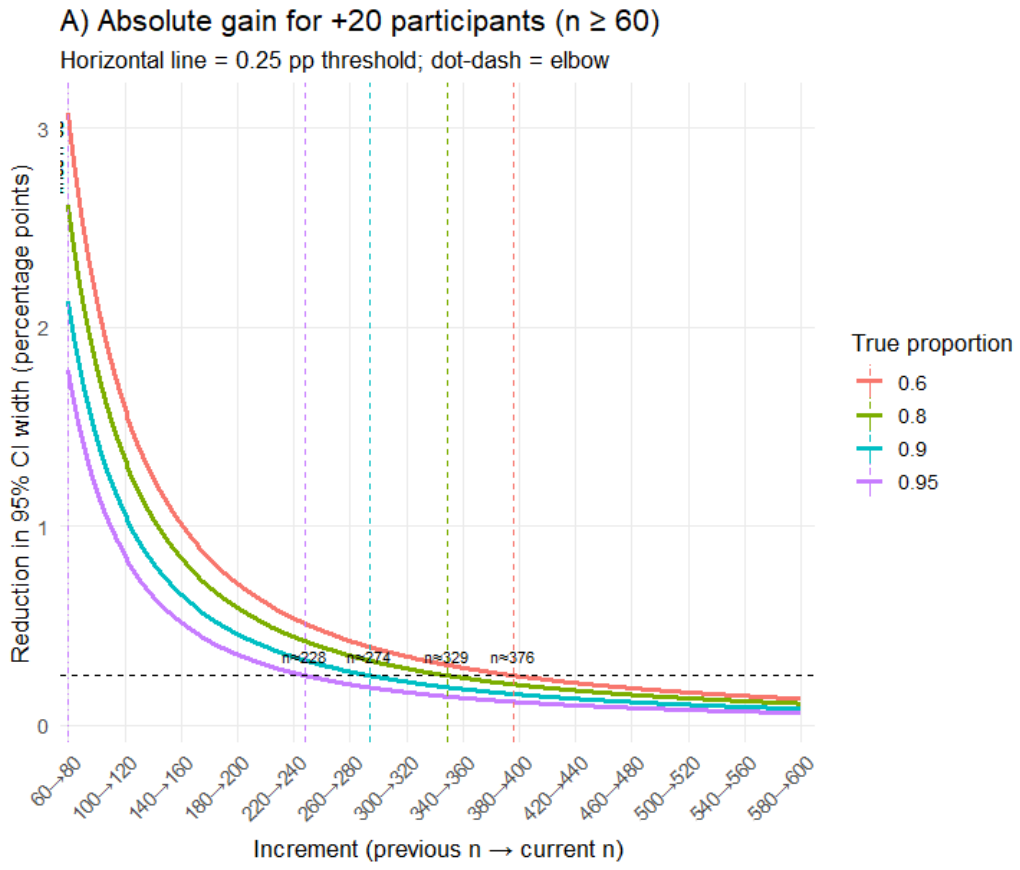
A4.3 Sample size

This is a start, and we will provide more description and guidance on this.

Figure x. Total width of a binomial confidence interval around a given proportion as a function of sample size.



The lines show the precision of estimates as a function of sample size; separate lines are plotted for different proportions (which may represent sensitivity or specificity). The y-axis shows total width of the 95% confidence interval for a proportion (which may represent sensitivity or specificity) for a given sample size, based on Wilson's score interval. The x-axis shows the required number of participants with tuberculosis (TB) to achieve a given level of precision for sensitivity (number of TB patients) or the number of participants without TB to achieve a given level of precision for specificity (number of non-TB patients). https://finddx.shinyapps.io/CI_plot/



For each 20-participant increase in sample size, the absolute reduction in 95% confidence-interval width declines steadily with increasing *n*. To provide some practical sense, one can observe that for each +20 in n (i.e. each 100 additional patients enrolled at a prevalence of 20%, when thinking of precision of the sensitivity estimate):

- Between ≈ 0-60, yields very large improvements in precision
- Between n ≈ 60-200, still yields very important improvements (1-3 % narrower Cis per +20).
- Between n ≈ 200-350, gains shrink to ~0.4-0.8 % per +20.
- Beyond n ≈ 230-380, gains drop below 0.25% per +20, marking the diminishing-returns region

The knee (or "elbow") in the precision curve was identified using a Kneedle-style curvature heuristic (Satopää et al., ICML 2011). This geometric criterion provides an objective, data-driven estimate of the transition to diminishing returns, which here coincides closely with the practical threshold of a 0.25-percentage-point reduction per +20 participants.

A4.4 Values

A variety of types of studies can be employed to inform the values domain, as shown in Fig. A2.1.

Fig. A2.x. Measures capturing people's views about health care outcomes

Table 1 Measures capturing people's views about healthcare outcomes

Utility measures		Non-utility, quantitative measures	Qualitative findings	
Direct techniques	Indirect techniques	Direct choice techniques	Interviews	Discussion groups
Standard gamble, time trade off, or visual analogue scales.	Pre-scored multi-attribute instruments [EQ-5D (euroQoL), the SF-6 health survey, or the health utility index (HUI-2 and HUI-3)]	Decision aids and direct choice studies. Surveys or questionnaires.	Structured, semi-structured, unstructured, or in-depth.	Focus groups.

Source: Selva et al. (2017) *(67)* (reproduced with permission of the authors).

Ideally, data from such studies would be compiled in a systematic review for the purposes of guideline development. Quantitative and qualitative methods provide different types of evidence, which can be complementary. Depending on the specific recommendation question at hand, one approach may be preferred over the other. Qualitative research, with its attention to the meanings that people assign to a phenomenon of interest and their understanding of that phenomenon can provide evidence on user values and preferences. Although qualitative evidence does not produce statistically generalizable results across a specific geography or population, it does provide analytical generalizability in terms of understanding the range of possible aspects that can explain a phenomenon of interest – in this case, the range of possible value considerations with regard to the intervention.

A4.5 Testing strategies

Using concurrent testing of different sample types offers a promising approach that considers the diagnostic testing barriers in people where a single test is unable to diagnose the condition. For example, for persons living with HIV, testing of sputum and urine during the same visit, when sputum can be produced, using LC-aNAATs and LF-LAM increases the likelihood of detecting TB with a rapid point-of-care result while also ensuring detection of rifampicin resistance.

Pooling of specimens can help in increasing test efficiency and cost effectiveness. By combining samples and testing them together, laboratories can test more individuals using fewer test reagents and less time, which is especially valuable in resource-limited or high-volume settings. This method enables broader population coverage, faster identification of TB cases, and optimized use of limited diagnostic resources. These strategies were used during COVID-19 pandemic and some calculators help assess the pool size based on disease prevalence and test accuracy.
https://bilder.shinyapps.io/PooledTesting/

A4.6 Estimands

Note: The idea of "estimands" has become increasingly important in treatment trials with specific regulatory guidance being used by EMA, FDA and other regulators and recognized in ICH (ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials).

It is essentially about being really clear and specific about how you define what you estimate or how you analyse data and in particular how you treat "intercurrent events". Different estimands can be

defined and address different questions. It is e.g. about what results you exclude when you have indeterminate results etc. (for reference standard, index test or comparators).

See also this related recent publication: Evans SR, Pennello G, Zhang S, Li Y, Wang Y, Cao Q, et al. Intention-to-diagnose and distinct research foci in diagnostic accuracy studies. The Lancet Infectious Diseases. 2025 Aug;25(8):e472–81.

Annex 5: Guidance on generating patient important outcome based evidence

TBD *[Note: This Annex will not be extensive or exhaustive on this topic. We started by adding a table with some of the literature that we think is particularly pertinent and helpful and which we think we could aim to summarize and refer to. Please feel free to suggest additions.]*

Table x. Overview of systematic reviews, conceptual and modelling papers relating to direct evidence of TB diagnostics on patient outcomes

Reference	Content
Nooy et al. Trade-Offs between Clinical Performance and Test Accessibility in Tuberculosis Diagnosis: A Multi-Country Modelling Approach for Target Product Profile Development. The Lancet Global Health 2024	Models trade-offs between accuracy and access
Haraka H et al. Impact of the diagnostic test Xpert MTB/RIF on patient outcomes for tuberculosis. The Cochrane database of systematic reviews. 2021	Systematic review and meta-analysis of eight randomized controlled trials (RCTs), and four before-and-after studies of Xpert MTB/RIF on patient outcomes
Ochodo EA et al. Variation in the observed effect of Xpert MTB/RIF testing for tuberculosis on mortality: A systematic review and analysis of trial design considerations. Wellcome Open Research. 2020	Systematic review and analysis of trial design considerations and exploration of heterogeneity
Ochodo EA et al. Improving the design of studies evaluating the impact of diagnostic tests for tuberculosis on health outcomes: a qualitative study of perspectives of diverse stakeholders. Wellcome Open Res. 2019	Perspectives of diverse stakeholders on improving the design of studies evaluating the impact of diagnostic tests for tuberculosis on health outcomes
Schumacher SG et al. The impact of Xpert MTB/RIF-do we have a final answer? The Lancet Global health. 2019	Discusses challenges and limitations of trials conducted to assess the effect of Xpert MTB/RIF on patient outcomes
Pai M et al. Surrogate endpoints in global health research: still searching for killer apps and silver bullets? BMJ global health. 2018	Discusses the role of surrogate outcomes and health systems when evaluating the impact of complex global health interventions on patient outcomes
Schumacher SG et al. Impact of Molecular Diagnostics for Tuberculosis on Patient-Important Outcomes: A Systematic Review of Study Methodologies. PLOS ONE. 2016	Systematic review and description of study methodologies, including RCTs, quasi-experimental studies and other non-randomized studies of interventions
Lawn SD et al. Effect of empirical treatment on outcomes of clinical trials of diagnostic assays for tuberculosis. The Lancet Infectious Diseases. 2015	Discusses the role of empirical treatment on the effect of TB diagnostics on patient outcomes
Sun AY et al. The impact of novel tests for tuberculosis depends on the diagnostic cascade. European Respiratory Journal. 2014	Models the role of factors in the diagnostic cascade on the effect of TB diagnostics on patient outcomes

Lin HH et al. The impact of new tuberculosis diagnostics on transmission: why context matters. Bull World Health Organ. 2012	Models the role of health systems context on the effect of TB diagnostics on patient outcomes

NOTE: We could also consider mentioning here somewhere some of the studies that were done using programmatic data such as e.g.

1. Hermans S, Caldwell J, Kaplan R, Cobelens F, Wood R. The impact of the roll-out of rapid molecular diagnostic testing for tuberculosis on empirical treatment in Cape Town, South Africa. Bull World Health Organ. 2017 Aug 1;95(8):554–63.
2. De Vos E, Westreich D, Scott L, Voss de Lima Y, Stevens W, Hayes C, et al. Estimating the effect of a rifampicin resistant tuberculosis diagnosis by the Xpert MTB/RIF assay on two-year mortality. PLOS Glob Public Health. 2023;3(9):e0001989.

List of useful references on quasi-experimental study designs and approaches to analysis

1. De Vocht F, Katikireddi SV, McQuire C, Tilling K, Hickman M, Craig P. Conceptualising natural and quasi experiments in public health. BMC Med Res Methodol. 2021 Dec;21(1):32.
2. Wing C, Simon K, Bello-Gomez RA. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. Annual Review of Public Health. 2018 Apr 1;39(1):453–69.
3. Feng S, Ganguli I, Lee Y, Poe J, Ryan A, Bilinski A. Difference-in-Differences for Health Policy and Practice: A Review of Modern Methods [Internet]. arXiv; 2024 [cited 2025 Jun 23]. Available from: http://arxiv.org/abs/2408.04617
4. Callaway B, Sant'Anna PHC. Difference-in-Differences with multiple time periods. Journal of Econometrics. 2021 Dec;225(2):200–30.
5. Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2016 Jun 9;dyw098.
6. Dimick JB, Ryan AM. Methods for Evaluating Changes in Health Care Policy: The Difference-in-Differences Approach. JAMA. 2014 Dec 10;312(22):2401–2.
7. Bärnighausen T, Røttingen JA, Rockers P, Shemilt I, Tugwell P. Quasi-experimental study designs series—paper 1: introduction: two historical lineages. Journal of Clinical Epidemiology. 2017 Sep;89:4–11.
8. Geldsetzer P, Fawzi W. Quasi-experimental study designs series—paper 2: complementary approaches to advancing global health knowledge. Journal of Clinical Epidemiology. 2017 Sep;89:12–6.
9. Frenk J, Gómez-Dantés O. Quasi-experimental study designs series—paper 3: systematic generation of evidence through public policy evaluation. Journal of Clinical Epidemiology. 2017 Sep;89:17–20.
10. Bärnighausen T, Tugwell P, Røttingen JA, Shemilt I, Rockers P, Geldsetzer P, et al. Quasi-experimental study designs series—paper 4: uses and value. Journal of Clinical Epidemiology. 2017 Sep;89:21–9.
11. Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series—paper 5: a checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels. Journal of Clinical Epidemiology. 2017 Sep;89:30–42.
12. Waddington H, Aloe AM, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, et al. Quasi-experimental study designs series—paper 6: risk of bias assessment. Journal of Clinical Epidemiology. 2017 Sep;89:43–52.
13. Bärnighausen T, Oldenburg C, Tugwell P, Bommer C, Ebert C, Barreto M, et al. Quasi-experimental study designs series—paper 7: assessing the assumptions. Journal of Clinical Epidemiology. 2017 Sep;89:53–66.
14. Glanville J, Eyers J, Jones AM, Shemilt I, Wang G, Johansen M, et al. Quasi-experimental study designs series—paper 8: identifying quasi-experimental studies to inform systematic reviews. Journal of Clinical Epidemiology. 2017 Sep;89:67–76.
15. Aloe AM, Becker BJ, Duvendack M, Valentine JC, Shemilt I, Waddington H. Quasi-experimental study designs series—paper 9: collecting data from quasi-experimental studies. Journal of Clinical Epidemiology. 2017 Sep;89:77–83.
16. Becker BJ, Aloe AM, Duvendack M, Stanley TD, Valentine JC, Fretheim A, et al. Quasi-experimental study designs series—paper 10: synthesizing evidence for effects collected from quasi-experimental studies presents surmountable challenges. Journal of Clinical Epidemiology. 2017 Sep;89:84–91.
17. Lavis JN, Bärnighausen T, El-Jardali F. Quasi-experimental study designs series—paper 11: supporting the production and use of health systems research syntheses that draw on quasi-experimental study designs. Journal of Clinical Epidemiology. 2017 Sep;89:92–7.
18. Rockers PC, Tugwell P, Røttingen JA, Bärnighausen T. Quasi-experimental study designs series—paper 13: realizing the full potential of quasi-experiments for health research. Journal of Clinical Epidemiology. 2017 Sep;89:106–10.
19. Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. International Journal of Epidemiology. 2018 Dec 1;47(6):2082–93.
20. Roth J, Sant'Anna PHC, Bilinski A, Poe J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. Journal of Econometrics. 2023 Aug;235(2):2218–44.

List of *some non-TB papers that have been influential or foundational as general methodological papers relating to patient important outcome based evidence of diagnostics:*

1. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ. 1986 Mar 15;134(6):587–94.
2. Lord SJ, Irwig L, Simes RJ. When Is Measuring Sensitivity and Specificity Sufficient To Evaluate a Diagnostic Test, and When Do We Need Randomized Trials? Ann Intern Med. 2006 Jun 6;144(11):850–5.
3. Lord SJ, Irwig L, Bossuyt PMM. Using the Principles of Randomized Controlled Trial Design to Guide Test Evaluation. Med Decis Making. 2009 Sep;29(5):E1–12.
4. Bossuyt PMM, Reitsma JB, Linnet K, Moons KGM. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem. 2012 Dec;58(12):1636–43.
5. Ruffano LF di, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. 2012 Feb 21 [cited 2025 May 20];
6. Staub LP, Dyer S, Lord SJ, Simes RJ. LINKING THE EVIDENCE: INTERMEDIATE OUTCOMES IN MEDICAL TEST ASSESSMENTS. International Journal of Technology Assessment in Health Care. 2012 Jan;28(1):52–8.
7. Staub LP, Lord SJ, Simes RJ, Dyer S, Houssami N, Chen RYM, et al. Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation. BMC Med Res Methodol. 2012 Feb 14;12:12.
8. Ferrante di Ruffano L, Dinnes J, Sitch AJ, Hyde C, Deeks JJ. Test-treatment RCTs are susceptible to bias: a review of the methodological quality of randomized trials that evaluate diagnostic tests. BMC Medical Research Methodology. 2017 Feb 24;17(1):35.
9. Ferrante di Ruffano L, Harris IM, Zhelev Z, Davenport C, Mallett S, Peters J, et al. Health technology assessment of diagnostic tests: a state of the art review of methods guidance from international organizations. Int J Technol Assess Health Care. 2023 Feb 21;39(1):e14.