
Claudia M. Denkinger,1,2 Samuel G. Schumacher,1 Christopher Gilpin,2 Alexei Korobitsyn,2 William A. Wells,3 Madhukar Pai,4 Mariska Leeflang,5 Karen R. Steingart,6 Michelle Bulterys,1,7 Holger Schünemann,8 Philippe Glaziou,9 and Karin Weyer7

1FIND, Geneva, Switzerland; 2World Health Organization, Geneva, Switzerland; 3USAID, Washington, District of Columbia; 4McGill International TB Centre and Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada; 5University of Amsterdam, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Netherlands; 6Liverpool School of Tropical Medicine, United Kingdom; 7University of Washington School of Public Health, Seattle; 8Department of Health Research Methods, Evidence and Impact and McMaster GRADE Centre, Hamilton, Canada; 9University Hospital Heidelberg, Division of Tropical Medicine, Centre of Infectious Diseases, Germany

Existing high-priority target product profiles (TPPs) of the World Health Organization (WHO) establish important needs for tuberculosis (TB) diagnostic development. Building on this earlier work, this guidance series aims to provide study guidance for performing accuracy studies of novel diagnostic products that may meet the 4 high-priority WHO TPPs and thus enable adequate evidence generation to inform a WHO evidence review process. Diagnostic accuracy studies represent a fundamental step in the validation of all tests. Unfortunately, such studies often have limitations in design, execution, and reporting, leading to low certainty of the evidence about true test performance, which can delay or impede policy and scale-up decisions.

This introductory paper outlines the following: (1) the purpose of this series of papers on study guidance; (2) WHO evidence needs and process for the development of policy guidelines for new TB diagnostic tests; and (3) study design considerations, ie, general diagnostic study considerations, intended use of test and role in the clinical pathway, choice of population and setting, index-test specific issues, suitable reference standard and comparators, study flow and specimen issues, and finally key issues beyond accuracy that should be considered. The other 4 papers in this series will provide more detailed guidance for each of the 4 WHO high-priority TPPs.

By increasing the clarity around the clinical evaluation needs for tests that have the potential to meet the TPP specifications, we hope to support harmonized evidence generation and enable the WHO review process towards meeting the WHO End TB Strategy targets for reducing the incidence and mortality associated with TB.

Keywords. diagnostics; target product profiles; TPPs; tuberculosis; WHO End TB strategy.
and technical challenges that hinder successful development of new TB diagnostics [8].

The pipelines to address the highest-priority needs (a standalone, nonspum-based point-of-care test or a point-of-care triage/rule-out test) are particularly meager. These tests would help to close the diagnostic gap and/or substantially reduce the cost of diagnosis, which has been identified as a major barrier to the uptake of existing tests.

Diagnostic accuracy studies represent a fundamental step in the validation of all tests. At the same time, diagnostic trials to generate evidence for global policy often have been perceived as a hurdle by industry given the costs and complexities associated with them. In addition, such studies often have limitations in design, execution, and reporting, leading to low certainty of the evidence about true test performance, which can further drive up cost, delay or impede policy, and scale-up decisions. This is a problem for diagnostic test accuracy studies in general [7–10] as well as for studies on TB in particular [11–14]. Recommendations for reporting and tools to assess risk of bias and applicability of study findings have been developed in response [15, 16]. However, the existing guidance only provides a general overview of design aspects to consider, without providing specific recommendations applicable to any particular disease or technology [15, 16]. Therefore, there is an urgent need for more specific guidance on study design to decrease risk of bias and avoidable heterogeneity.

This series of guidance papers aims to highlight the evidence needs and provide guidance on the design of diagnostic test accuracy studies of tests that meet the high-priority TPPs (Box 1). The TPPs contain a wide range of requirements for test solutions that need to be considered when designing or evaluating new products. The diagnostic accuracy of a test is arguably the most fundamental attribute that needs to be established to allow assessment of its potential value; without good accuracy, other outcomes such as clinical impact or cost-effectiveness cannot be determined. It is also more challenging to evaluate accuracy reliably than other important test attributes such as test operational characteristics. Thus, this guidance series focuses on this important aspect. We do not address how other product characteristics should be measured, although we do provide references to existing guidance for other aspects where available. A TPP for new tests for latent/subclinical TB has been published separately alongside guidance for clinical studies to assess their performance [17].

**WORLD HEALTH ORGANIZATION PROCESS FOR THE DEVELOPMENT OF POLICY GUIDELINES FOR NEW TUBERCULOSIS DIAGNOSTIC TESTS**

Within WHO, the review of data on new TB diagnostic tests is performed by WHO's Global TB Programme [9]. The WHO Prequalification process does not yet apply to TB given that most TB tests have single-source manufacturers using unique technologies. This might change in the future, particularly as more tests meeting the same TPP come to the market.

There are 2 principal ways in which WHO approaches the review of data on TB diagnostics: (1) for the review of a truly new diagnostic technology or a novel or expanded intended use, WHO convenes a Guideline Development Group (GDG) to evaluate a body of evidence using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach; and (2) for the revision of existing guidance, either a GDG is convened or a Technical Expert Group (TEG) consultation is used to assess technical documentation [10].

The outcome of a GDG meeting is a new or updated WHO Guideline, whereas the outcome of a TEG consultation is a WHO Technical Report. Examples for diagnostic tests that have recently undergone the 2 different pathways are as follows: first- and second-line Line Probe Assays were reviewed in a GDG in 2016, which resulted in a diagnostic guideline that was issued in parallel with the guideline for the use of short-course, multidrug-resistant (MDR) therapy [11–13]. The Xpert MTB/RIF Ultra, a next-generation test after the Xpert MTB/RIF, was assessed for equivalent performance in a TEG consultation in 2017. A formal GDG process will be held in 2019 to refine and update the current Xpert guidelines [14]. To include a diagnostic test in the WHO's List of Essential In Vitro Diagnostics (EDL), it must have been recommended for use by a WHO GDG. The WHO updates the EDL on an annual basis to include new diagnostics that have been assessed via this robust evaluation process [15].

The Foundation for Innovative New Diagnostics (FIND), a WHO collaborating center that evaluates new diagnostic technologies, updates the WHO Global TB Programme on the TB diagnostics pipeline on a regular basis [8]. If new versions of the WHO-recommended assays are available, WHO needs to be provided with data that demonstrate the equivalence of performance by the manufacturer. All diagnostic policy guidelines are reviewed as new evidence becomes available, and these are normally updated every 3 to 5 years.

In advance of a WHO GDG or TEG meeting, all available evidence on a product is identified and synthesized in a systematic review (and meta-analysis, if appropriate). Alternatively, it is possible for sufficient evidence to be provided by a single multicenter study of high quality that is conducted with the

---

**Box 1: Papers Included in This Guidance Series**

- **Paper 1.** Introduction to TB diagnostics study guidance series.
- **Paper 2.** Study guidance: Smear replacement tests.
- **Paper 3.** Study guidance: Biomarker-based tests.
- **Paper 4.** Study guidance: Triage tests.
- **Paper 5.** Study guidance: Centralized DST and sequencing.
specific objective to assess a particular technology in sites representative of the global TB epidemic, as done for the Xpert MTB/RIF Ultra [14].

The data necessary for WHO review (Table 1) need to come from the following: (1) an analytical validation; and (2) a clinical validation compared with a reference method [16] (online supplement: Glossary). Analytical validation refers to measuring accuracy, precision, and reproducibility of the test in contrived specimens or panels [18]. This work often confirms assessments already done by the diagnostic manufacturers, in the hands of an independent evaluator, and may allow researchers to reduce the sample size needed in prospective clinical studies that assess clinical validity. For example, testing strains that harbor key drug-resistance conferring mutations can be an efficient way to assess performance of drug-resistance assays and can reduce (but not eliminate) the need for directly testing clinical specimens. Well characterized frozen specimens may also be used to complement prospective clinical studies, recognizing the fact that the frozen samples have the limitation of the altered matrix. Clinical validation refers to a prospective clinical study that assesses the accuracy with which a test identifies a patient’s clinical status [19].

Although data on analytical and clinical validation will always be necessary, the evaluation of clinical utility (eg, impact of a diagnostic test on patient important outcomes such as time to treatment initiation or mortality) in demonstration studies is not always performed in advance of a first WHO evidence review [19, 20]. While such demonstration studies certainly add important information for implementation considerations, they are often only considered after a first WHO review in order not to delay introduction of an assay. Furthermore, clinical utility is best assessed if a test is used for clinical care, which is not possible before a regulatory approval. Whether or not a demonstration study is necessary in advance of a first WHO review is decided by the WHO, but evidence of clinical utility is typically most critical for “disruptive technologies” that lead to important changes in clinical pathways. For example, implementing a triage test meeting the TPP characteristics would dramatically change algorithms and be used in settings where currently no diagnostic testing for TB is performed. Thus, an accuracy study is unlikely to paint the whole picture necessary to guide introduction of such an assay. In contrast, a new molecular TB detection assay replacing an existing molecular TB detection assay is unlikely to need a demonstration study in addition to an accuracy study. Obviously, the need for a demonstration study has financial implications to companies. That being said, studies for innovative technologies are often grant funded.

Key data needs (Table 1) for a GDG review align with those of stringent regulatory authorities (SRAs), such as the US Food and Drug Administration, European Commission CE marking under the new directive, and the Global Harmonization Task Force [21–23]. Thus, much of the evidence generated for a WHO policy development process can be used in parallel for a CE submission or similar regulatory process, to avoid delays in guideline development and regulatory approval. A WHO review, although assessing similar data as the SRAs, also considers the specific patient population targeted by the guideline, the level of the health system for implementation, and the needs and challenges in high-burden countries with varying epidemiology of HIV-associated TB and MDR-TB. A WHO review also includes consideration of patient values and preferences, resource use, feasibility, acceptability, and equity [24].

The evidence generated for a GDG meeting will be subject to a GRADE assessment, which rates the certainty (also called “quality”) of the scientific evidence in diagnostic trials that have been synthesized in systematic reviews and allow for the development of evidence-based recommendations in guidelines with a process that is fully documented and transparent [25]. The GRADE’s 4 categories of certainty of evidence (very low, low, moderate, high) imply increasing confidence in estimates of the effect of a diagnostic test or strategy on proportions of true and false positives and true and false negatives. Within the GRADE framework, evidence is graded based on study design, risk of

<table>
<thead>
<tr>
<th>Table 1. Types of Data Needed and Key Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Needs</strong></td>
</tr>
<tr>
<td>Analytical validity</td>
</tr>
<tr>
<td>Clinical validity</td>
</tr>
<tr>
<td>Clinical utility (patient impact)</td>
</tr>
<tr>
<td>Epidemiologic utility (population impact)</td>
</tr>
<tr>
<td>Economic outcomes</td>
</tr>
<tr>
<td>Operational/implementation aspects</td>
</tr>
<tr>
<td>Values and preferences</td>
</tr>
</tbody>
</table>

Introduction to Study Design Guidance Series • JID 2019:220 (Suppl 3) • 593
bias, indirectness, imprecision, inconsistency, indirectness, and other considerations, such as publication bias [26]. The often indirect impact of diagnostic tests on patient-important outcomes (such as mortality) is acknowledged in this framework and in the stakeholder community at large [27, 28].

**STUDY DESIGN CONSIDERATIONS**

Diagnostic test accuracy studies need to conform with agreed-upon principles that consider ethics, design, conduct, and reporting (Table 2) [29]. Specifically, studies need to be designed with consideration of Good Clinical Practice to ensure that the rights, safety, and well-being of research subjects are protected and respected, consistent with the principles enunciated in the Declaration of Helsinki and other internationally recognized ethical guidelines [30, 31]. The study design should minimize the risk of bias across the 4 key domains identified within the QUADAS-2 tool: patient selection, index test, reference standard, and flow and timing [32]. The reporting of the diagnostic accuracy studies that assess clinical validity should follow the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidance to ensure that all essential information about the study is provided [33]. Systematic reviews should (1) follow standard methods such as those described by Cochrane [34] and (2) be reported using the preferred reporting items for systematic review and meta-analysis (PRISMA) guideline [35, 36]. Useful guidance is also available from the Agency for Healthcare Research and Quality [34, 37].

This guidance series will focus on the trial needs to address analytical and clinical validity for the 4 high-priority TPPs that can substantially improve the currently existing TB diagnostic cascade (Box 1). The structure of the papers on the guidance for studies for tests meeting the individual TPPs is described here. The diagnostic accuracy studies for the first 3 TPPs (papers 2–4) will focus on the following: (1) the intended use of the test and its implication for the study design; (2) general study design considerations; (3) choice of population and setting; (4) issues pertaining to the index test (the test under investigation) itself; (5) reference standard and comparators; (5) flow and specimen issues; and (6) key issues beyond accuracy that should be considered. General considerations and definitions applicable to all tests are addressed in more detail in this paper.

**Intended Use**

Based on the STARD reporting and GRADE approach to rating certainty and developing recommendations, the “intended use of the test” is defined as a combination of the use case (ie, whether the index test is used for diagnosis, screening, staging, prediction, or other reasons) and which populations, clinical settings, and interventions the test intends to target [20, 33, 38]. It is crucial to define the objective intent of the test in a clear and comprehensive statement that encompasses these elements. These conditions will ultimately drive the clinical study design to assess the test’s performance (how well it achieves its intended purpose) and inform the review process for policy [18].

**General Study Design Considerations**

The diagnostic accuracy studies for the first 3 TPPs (papers 2–4) should be a cross-sectional study of either a consecutive series or a random sample of unselected patients who require evaluation for TB. For a study to assess DST, the study design depends on whether the test is intended to be a follow-on test after *Mycobacterium tuberculosis* (MTB) has already been identified, or whether it also aims to be a simultaneous test of TB detection and DST. This will also define the study population. Sample size and the role of analytical data and banked specimens will vary by the diagnostic test evaluated and are addressed in the subsequent papers.

**Choice of Population and Setting**

For studies evaluating tests that aim to identify TB disease, the initial study population is adults with respiratory symptoms suggestive of TB. In peripheral settings of care, patients might present with early forms of disease if access to care is readily available. This might have an impact on the test performance (accuracy) and also on the prevalence and predictive values of a test (relevant for GRADE). However, clinical validity studies as described here might not provide a full picture of this complexity because the studies will often need to be conducted in settings in higher levels of care to provide a highly controlled environment for the clinical study and reference standard.

**Table 2. Sources for Guidance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Information Provided</th>
<th>Reference No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICH and GCP</td>
<td>General considerations for ethics and conduct</td>
<td>[31]</td>
</tr>
<tr>
<td>QUADAS-2</td>
<td>Tool for assessment of risk of bias (and for planning to prevent it)</td>
<td>[32]</td>
</tr>
<tr>
<td>DEEP</td>
<td>Guidance on conduct of infectious disease diagnostics studies</td>
<td>[29]</td>
</tr>
<tr>
<td>STARD</td>
<td>Reporting guidance</td>
<td>[33]</td>
</tr>
<tr>
<td>Cochrane Handbook, AHRQ Methods Guide, PRISMA-P</td>
<td>Guidance for conducting systematic reviews of diagnostic test accuracy</td>
<td>[33, 34, 36]</td>
</tr>
<tr>
<td>PRISMA and PRISMA-DTA</td>
<td>Guidance for reporting systematic reviews of diagnostic test accuracy</td>
<td>[35, 46]</td>
</tr>
<tr>
<td>GRADE and EtD</td>
<td>Evidence review for policy making</td>
<td>[20, 24–27]</td>
</tr>
</tbody>
</table>

Abbreviations: AHRQ, US Department of Health and Human Services Agency for Healthcare Research and Quality; DEEP, Diagnostics Expert Evaluation Panel; DTA, Diagnostic Test Accuracy; EtD, GRADE Evidence to Decision; GCP, good clinical research practice; GRADE, Grading of Recommendations Assessment, Development and Evaluation; ICH, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; PRISMA, preferred reporting items for systematic review and meta-analysis; QUADAS, Quality Assessment of Diagnostic Accuracy Studies; STARD, Standards for Reporting of Diagnostic Accuracy Studies.
because otherwise the data generated on accuracy will be difficult to interpret. Although demonstration studies will be able to adequately reflect upon these differences in patient populations, they will typically not have the reference standard to highlight the differences.

Given the large percentage of patients with TB/HIV coinfection that forms part of the global TB epidemic, patients living with HIV should be included as part of the study for all test types and should be analyzed as a subgroup. For tests that assess samples other than sputum (triage test and nonsputum biomarker-based detection test), patients who have symptoms suggestive of extrapulmonary TB should also be considered, provided a sufficient reference standard (likely a composite reference standard based on microbiological tests, radiology, pathology, and clinical characteristics) can be established. The same applies to pediatric TB cases, where the additional complexity with regards to the reference standard often limits the inclusion into the clinical validity study (also see more detailed discussion in “Paper 3”). Nevertheless, all efforts should be made to include children, because they are a vulnerable and neglected group [25].

When a WHO evaluation process relies on evidence from only 1 large multicenter study for clinical validity, it is particularly important that the study assess the novel technology using a standardized protocol and reference standard, in sites representative of the global TB epidemic, ideally in countries/settings that are archetypal for the region, to help support broad generalizability of the findings. Several such multicenter studies have been conducted and used for WHO policy development [13, 14].

**Index Test**

Study design considerations with regards to the index test are highly dependent on the biomarker and test platform (see specifics in following papers).

**Reference Standard and Comparators**

A microbiological reference standard remains the best available reference standard for TB. At a minimum, a single liquid sputum culture (with speciation) should be considered. Optimally, 2 liquid cultures on 2 separate samples, provided on 2 separate days, would be done for all patients. This is particularly important because the sensitivity of novel tests is approaching that of culture, and false-positive index-test results need to be ruled out. However, the pitfalls of culture as a reference standard are numerous and should be considered carefully: (1) culturing methodology in itself is highly complex and prone to variability and error (eg, over-decontamination), which can result in misclassification; (2) conceivably, biomarkers that detect nonpathogen markers might detect earlier stages of disease that are not yet culture positive or extrapulmonary TB that is not captured by a sputum culture; (3) accuracy estimates can only be compared between studies if the number and type (liquid versus solid) of cultures are the same. The same considerations apply to the subanalysis of data, ie, data can only be compared if the subgroups have been defined the same way. For example, estimates of sensitivity by smear status can differ widely depending on whether one or multiple smears, or Ziehl-Neelsen light microscopy or fluorescence microscopy, have been used to define smear status [14]. Likewise, an analysis of pediatric TB can vary depending on how the composite reference standard has been defined. Before starting a study, researchers should also
carefully consider any additional work-up needed (and redefine it in the operating procedures and analysis plan) to resolve discrepant results (eg, sequencing of amplicons or deoxyribonucleic acid extracts of a molecular test or repeat testing of left over samples). Again, this is getting increasingly important because novel tests are reaching close to the sensitivity estimates of the reference standard or the reference standard itself is getting questioned (as is the case for phenotypic DST) [39].

Composite reference standards, including additional testing from nonspuam samples and/or clinical diagnosis, may need to be considered particularly for nonspuam-based tests for MTB detection. Statistical techniques (eg, latent class analyses) can also be considered to better handle analysis in the context of an imperfect reference standard [40]. A sequencing reference standard (alone or in combination with a phenotypic reference standard depending on the drug evaluated) should be considered for studies on DST.

Using WHO-recommended tests as comparator tests (eg, Xpert MTB/RIF in the assessment of other molecular tests used to MTB detection) allows for benchmarking against a test with the same intended use for which a large evidence base exists. This can also protect against the risk of spectrum bias. Note that comparators should not be part of the reference standard although their results can aid interpretation of results that are discordant between index test and reference standard.

Flow and Specimen Issues
Depending on the index test and planned comparator tests, the study flow needs special consideration. Ideally, the index test, comparator test, and reference standard should be performed on the same specimen. However, this might lead to high specimen volume requirements, which can result in biasing the study population (eg, patients with paucibacillary disease are unlikely to produce high-volume samples).

Key Issues Beyond Accuracy
Although accuracy is a key piece of evidence, the assessment of operational characteristics of a diagnostic test (eg, the time taken to perform the test, its technical simplicity or ease of use, and user acceptability), the connectivity solutions, and training materials are important as well [24, 41]. These assessments should be part of an accuracy study, although it must be acknowledged that the users in the context of an accuracy evaluation will likely be more trained and experienced, and there will be fewer challenges on connectivity than in real-world implementation. Thus, such an assessment should be repeated as part of a demonstration study. Other aspects such as feasibility and equity also form part of the criteria for formulating recommendations according to the GRADE Evidence to Decision Frameworks [20]. The effect of tests on intermediate outcomes that imply impact on patient outcomes (eg, reduced time to diagnosis and treatment) also needs to be considered. However, this is better performed in the context of a demonstration study where the test is used for patient care (ie, after regulatory or policy approval for clinical use), which is often not the case in the context of initial accuracy evaluations. Economic analyses should also be part of the assessment to inform the WHO policy process [42, 43].

Although a WHO recommendation on a diagnostic test carries a lot of weight and enables procurement of a product with funding from the Global Fund, translation of global policy into actionable implementation plans at the country level often requires additional in country studies. Therefore, specific country and donor engagement plans are required to ensure translation of global policy into actionable implementation plans [44].

CONCLUSIONS
This introduction paper sets the stage for papers 2–5 of this series of TPP study design guidance documents. Study design considerations differ greatly depending on the TPPs, and these considerations are addressed in detail in the subsequent papers. The series aims to increase clarity around the clinical evaluation needs for tests that have the potential to meet the TPP specifications published by the WHO [5, 7], with the goal to facilitate the evaluation of such tests and move the field forward towards meeting the WHO End TB Strategy targets [45].

Supplementary Data
Supplementary materials are available at The Journal of Infectious Diseases online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Supplement sponsorship. This supplement is sponsored by FIND (Foundation for Innovative New Diagnostics) and was made possible through the generous support of the Governments of the United Kingdom, the Netherlands, Germany and Australia.

Disclaimer. The views and opinions expressed in this article are those of the author and not necessarily the views and opinions of the US Agency for International Development (USAID).

Potential conflicts of interest. W. A. W. is employed by USAID (Washington DC). All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References


