Guidance for Studies Evaluating the Accuracy of Rapid Tuberculosis Drug-Susceptibility Tests

Sophia B. Georghiou,[1] Samuel G. Schumacher,[1] Timothy C. Rodwell,[1] Rebecca E. Colman, [1] Paolo Miotto,[2] Christopher Gilpin,[3] Nazir Ismail,[4,8,9] Camilla Rodrigues,[5] Rob Warren, [6] Karin Weyer,[3] Matteo Zignol,[3] Sonia Arafah,[1] Daniela Maria Cirillo,[2] and Claudia M. Denkinger[1,7]

[1]FIND, Geneva, Switzerland;

[2]IRCCS San Raffaele Scientific Institute, Milan, Italy;

[3]World Health Organization, Geneva, Switzerland;

[4]University of Pretoria, South Africa;

[5]Hinduja Hospital and Medical Research Centre, Mumbai, India;

[6]SAMRC Centre for Tuberculosis Research, Stellenbosch University, Tygerberg, South Africa;

[7]University of Heidelberg, Centre of Infectious Diseases, Germany;

[8]National Institute for Communicable Diseases, National Health Laboratory Service, Johannesburg, South Africa;

[9]University Hospital Heidelberg, Division of Tropical Medicine, Centre of Infectious Diseases, Germany

Abstract

The development and implementation of rapid molecular diagnostics for tuberculosis (TB) drug-susceptibility testing is critical to inform treatment of patients and to prevent the emergence and spread of resistance. Optimal trial planning for existing tests and those in development will be critical to rapidly gather the evidence necessary to inform World Health Organization review and to support potential policy recommendations. The evidence necessary includes an assessment of the performance for TB and resistance detection as well as an assessment of the operational characteristics of these platforms. The performance assessment should include analytical studies to confirm the limit of detection and assay ability to detect mutations conferring resistance across globally representative strains. The analytical evaluation is typically followed by multisite clinical evaluation studies to confirm diagnostic performance in sites and populations of intended use. This paper summarizes the considerations for the design of these analytical and clinical studies.

Keywords: diagnostics, drug-susceptibility testing, target product profile, tuberculosis, WHO End TB Strategy

The rapid and accurate diagnosis of tuberculosis (TB) and determination of drug susceptibility is critical for patient treatment and to prevent the emergence and spread of resistant strains. Globally, in 2017, less than one third of new TB patients received drug-susceptibility testing (DST) for rifampicin (RIF), one of the most important first-line drugs [1]. This leads to the undertreatment of drug-resistant TB (DR-TB), further amplification and transmission of resistance, and associated mortality [2]. Modeled data predict a rising incidence of multidrug-resistant (MDR)-TB, with an increase of 9%–33% in the Philippines, India, and Russia alone [3, 4]. In 2017, only 50% of those diagnosed with RIF-resistant TB in 2017 had second-line DST performed, even when MDR-TB was suspected [1]. The lack of information on second-line drug susceptibility, especially for TB DST target product profile (TPP) priority compounds such as the fluoroquinolones (FQs) [5], can lead to catastrophic outcomes for patients, an increased burden on health-systems, and the transmission of resistant TB. Without addressing these key issues, the 2030 Sustainable Development Goal of ending the TB epidemic will not be reached.

In view of the limitations of conventional phenotypic methods, the development of rapid molecular diagnostics for TB DST has become a research and development priority [6]. Although the rollout and uptake of novel DR-TB diagnostics such as the Xpert MTB/RIF assay (Cepheid, Sunnyvale, CA) has increased the number of TB and DR-TB cases detected and notified [1, 7], important diagnosis and treatment gaps remain. In particular, there is a pressing need for rapid molecular DSTs that detect resistance to a wider array of drug compounds, including those prioritized in the TPP [5]. Several novel assays have been developed in line with the existing TPP [5], and some have already demonstrated promising performance for *Mycobacterium tuberculosis* complex (MTBC) and drug resistance detection in early studies [8–13]. However, it is rare that studies of these technologies adequately inform World Health Organization (WHO) review and support potential policy recommendations. For example, these studies (1) often fail to include well characterized comparator assays, (2) do not test an adequate selection of resistance mutations in strains of wide geographic variance, (3) do not include adequate sample size to achieve diagnostic accuracy precision targets, or (4) use a sample flow that does not allow for robust comparisons between the index test, reference test, and comparators.

In this article, we define standards for the generation of evidence for TB DST solutions to ensure that analytical and clinical evaluations answer key questions to enable comprehensive technology review. We summarize our recommendations in Table 1.

Table 1. Overview of Recommendations for Study Design

Topic	Recommendation
Index test	• Consider whether test is a reflex test to a TB-positive result or whether test also detects MTBC in addition to resistance when designing studies
	• Consider specifics of the index test under investigation (eg, throughput, polyvalency, connectivity) in study design
General study design considerations	• Initial laboratory analytical studies should confirm assay accuracy for MTBC and resistance detection (a broader range of mutations can be covered than in a clinical study)
	• Data on assay inclusivity and exclusivity should be made available by manufacturers, and data on resistance detection in heteroresistant samples is desirable
	• Clinical evaluation studies should evaluate the diagnostic performance of TB drug susceptibility tests on patient specimens in settings of intended use
Population and setting	• Ideally, at least 3 sites from 3 different WHO regions should be selected that reflect the TB and DR-TB epidemic in high-burden countries
	• Perform testing in the intended setting of use (ie, at central-level reference laboratories for high-throughput tests, and primary healthcare centers for lower-throughput or single-use tests)
	• Sample size targets should be set with consideration for TPP-defined target performance estimates for each drug to which the assay claims to detect resistance. Ideally, the width of confidence intervals should be ≤5% for specificity estimates and ≤10% for sensitivity estimates.
Reference standard and comparators	• Given the need for a comprehensive picture of drug resistance and the resolution of discrepancies, the use of a composite reference standard that combines genotypic sequencing information and phenotypic susceptibility testing results is recommended
	• Include WHO-approved tests (ie, the WHO-endorsed line probe assays and the Xpert MTB/RIF or Ultra assays) as comparators in both analytical and clinical studies
Flow and specimen issues	• Ideally, the index, reference and comparator tests should all be performed on the same sample to ensure comparability of results
	• Alternatively, for sputum-based tests, the index test may be performed twice: once on the direct sputum sample and once on the cultured isolate along with the reference and comparator tests
Key issues beyond accuracy	• Various assay technical and operational performance parameters that should be evaluated include time-to-result, invalid and indeterminate rates, and other factors (Table 2)

Abbreviations: DR-TB, drug-resistant tuberculosis; MTBC, *Mycobacterium tuberculosis* complex; RIF, rifampicin; TB, tuberculosis; TPP, target product profile; WHO, World Health Organization.

INDEX TEST: INTENDED USE AND TUBERCULOSIS DRUG-SUSCEPTIBILITY TEST ASSAY PIPELINE

The intended use of TB DST assays is to detect anti-TB drug resistance directly from clinical samples. These assays may be used as an up-front test and include MTBC detection or only as a reflex test to a positive result from MTBC-detection assays. High-throughput DST assays are typically used at central-level reference laboratories, including level-3 referral laboratories or level-2 district hospitals, whereas lower-throughput DST assays may be implemented at level-1 centers.

A few recently developed assays currently on the pathway to WHO approval and aimed for use in centralized laboratories include the Abbott RealTi*m*e MTB and MTB RIF/INH assays (Abbott, North Chicago, IL), the Roche COBAS MTB and MTB-RIF/INH assay (Roche Diagnostics, Basel, Switzerland), the Hain FluoroType MTBDR assay (Hain Lifescience GmbH, Nehren, Germany), and the BD MAX MDR-TB assay (Becton Dickinson, Franklin Lakes, NJ) [8–11]. The Abbott RealTi*m*e MTB assay can diagnose MTBC in 94 samples, with positive specimens reflexed to the RIF/INH assay for MDR-TB diagnosis within 10.5 hours [8, 14]. The Roche COBAS MTB assay also uses real-time polymerase chain reaction (PCR) for MTBC detection and can generate results for 96 tests in one 3.5-hour run, with positive specimens reflexed to the RIF/INH assay for MDR-TB diagnosis an additional 3.5 hours later [15]. The Hain FluoroType MTBDR assay relies upon LATE-PCR amplification and light on/lights off chemistry to detect MTBC and isoniazid (INH) and RIF resistance for 94 samples within 4 hours [16]. A new assay capable of FQ and second-line injectable (SLI) resistance detection using this same technology is currently under development [13]. The BD Max MDR-TB assay is another real-time PCR assay that can be run on the BD MAX System to detect MTBC and INH and RIF resistance for 22 sputum samples in 4 hours [11]. Targeted next-generation sequencing assays will also be an option for versatile centralized TB DST in

the near-term [17]. One novel test currently in the pipeline for expanded DST in decentralized settings is the Xpert MTB/XDR assay (Cepheid), which can be run on the GeneXpert platform for INH, FQ, and SLI resistance detection [12]. The Molbio Truenat assay (MolBio Diagnostics Pvt Ltd, Goa, India) that enables MTBC and RIF resistance detection was recently approved for use in India and is undergoing trials for WHO review [18]. An overview of additional DST assays in development, or undergoing validation or regulatory approval, is available through FIND's (Foundation for Innovative New Diagnostics) diagnostic pipeline tracker [19].

These TB DST assays all claim high sensitivity and specificity for resistance detection in TB clinical samples. Many have additional characteristics that are of added value, including polyvalency, ie, detection and differentiation of nontuberculous mycobacteria (NTM) and viruses (eg, human immunodeficiency virus) on the same platform, and/or platform connectivity to facilitate results reporting and sharing (Table 2) [5]. Manufacturers should provide data on these additional characteristics whenever possible. Given the recent update to WHO DR-TB treatment guidelines [20], it is likely that novel assays will soon be developed that also test for resistance to newer drugs (eg, bedaquiline and/or linezolid) once the molecular basis of resistance to these drugs is well defined [21]. Similar study design considerations will apply to these assays.

Table 2. Assay Operational Characteristics[a]

USE
• Total hands-on time required for an assay run
• Total number of steps required from sample reception to result output (and time requirement for substeps, eg, DNA extraction)
• Time to first result (from sample reception to result output)
• Ease of use and user appraisal over time (comparison at 2 different time points: after training and at the end of study)
• Training needs (eg, number of runs required to achieve proficiency during training, prior technical expertise needs, time to pass proficiency testing)
ASSAY
• Batch efficiency (minimum, maximum)
• Assay operating temperatures (minimum, maximum)
• Sample volume requirements (minimum, maximum), including by sample type
• Reagent storage requirements, including reusability between runs, shelf life
• Reagent use/waste (eg, volume that reagents come in, volume of reagents required to run a test, how long the reagents can be stored for after opening/thawing, etc) and disposal methods
• Lot-to-lot variability (based on information provided by manufacturer) or quality issues of reagents as per Incoming Quality Check results
• Shipping and storage requirements (eg, reagents require cold chain for shipment)
• Reagent stability dating (eg, average expiry date from date of shipment, expiry date from first use)
INSTRUMENT
• Instrument stability
• Voltage and power requirements
• Polyvalency: other tests available on the instrument
• Random access
• Minimal infrastructure requirements for testing, including square footage, storage, and waste disposal
• Instrument failure rate (both device-related and user-related errors)
• DNA contamination potential (based on information provided by manufacturer) and actual number of events (eg, if apparent, based on series of false-positive index test results after processing of a positive specimen)
• Maintenance and customer support needs, based on issues reported by sites and periodic assessments, as well as recommended service intervals
• Integration with Laboratory Information Systems, remote access for technical support and device monitoring, key performance indicators

[a]Assay operational characteristics may be captured during clinical trials, as part of separate studies, or may be obtained from information provided by manufacturers and taken into consideration when judging the suitability of a new assay or platform for different environments.

GENERAL STUDY DESIGN CONSIDERATIONS

Two types of studies should be considered to assess accuracy and reliability of novel assays for DST to support WHO and country policy making processes: (1) analytical studies in laboratory settings to confirm assay limit of detection for MTBC and resistance detection; and (2) clinical evaluation studies to confirm diagnostic performance on clinical samples collected from a consecutive series or random sample of unselected patients requiring evaluation for TB and DR-TB in sites of intended use (Appendix 1: Glossary). Together, these studies can provide representative data on assay and instrument performance for use in TB high-burden countries and utility across different clinical settings and populations (Table 3). Ease-of-use assessments and other assessments (see Table 1 in the paper 1 by Denkinger et al) are also a requirement but will not be addressed herein.

Table 3. Use of Evidence From the Analytical and Clinical Evaluation Studies to Address Objectives[a]

		Analytical Study		
Assessed	Characteristic	LoD Panel	Resistance Panel	Clinical Study
MTB-detection	Sensitivity[b]	○		●
	Specificity[c]			●
	Direct vs pellet			●
	Head-to-head (Xpert)[d]	○		●
Resistance-detection	Sensitivity[e]		●	○
	Specificity		○	●
	Direct vs pellet			○
	Head-to-head (LPAs)[f]		●	
Operational characteristics	Various			○

Abbreviations: LoD, limit of detection; LPA, line probe assay; MTB, *Mycobacterium tuberculosis*; RIF, rifampicin; TB, tuberculosis.

[a]Filled circles (●, main evidence to address given trial aim; open circles (○), evidence supportive or confirmative of main evidence.

[b]Sensitivity estimates from the clinical trial provide the main evidence and are generally supported by the comparative limit of detection assessment.

[c]Specificity estimates may be mainly based on clinical trial data and are generally supported by data on exclusivity testing provided by the assay developer.

[d]Direct head-to-head comparison of assay to comparators will further increase confidence in the sensitivity estimate (eg, including Xpert as a comparator for the detection of paucibacillary TB, due to the large amount of data available for Xpert MTB/RIF).

[e]The main evidence for sensitivity for detection of resistance-conferring mutations will come from phase 1 because the number of patients with resistant TB as well as the variety of mutations will be limited in a clinical study.

[f]Direct head-to-head comparison to comparators (eg, Hain line probe assay) will further increase confidence in the sensitivity estimate for resistance detection.

ANALYTICAL STUDIES

Independent analytical studies will complement and confirm what manufacturers produce for their assay verification. These studies should be conducted in laboratory settings on the design-locked assay to assess MTBC and resistance detection limits, analytical sensitivity and specificity, inclusivity and exclusivity, and heteroresistance detection against a range of well characterized samples.

Limit of Detection

Given that prospective clinical studies suffer from tested patient population variability (over time, in different geographies, or in different catchment areas), analytical studies of limit of detection (LoD) performed on a standardized panel and against well characterized comparator assays can provide an LoD that can be compared across assays. Ideally, LoD testing should be conducted in a validated sputum matrix to gather sufficient data to fit a Probit curve and estimate the LoD with a 95% confidence interval. Given the extensive data available on Xpert MTB/RIF and Ultra, the inclusion of either assay as a direct comparator is recommended to enable benchmarking. Results should confirm that the assay LoD for MTBC is equivalent or superior to at least that reported for Xpert MTB/RIF [17], in line with TPP

criteria [5], although reduced sensitivity may be acceptable if other assay characteristics would substantially improve availability and access [22]. Limit of detection testing against Ultra may be particularly useful for highly sensitive DST assays to inform placement in testing algorithms (eg, use as either an up-front or reflex test).

The LoD for DST resistance targets should also be confirmed, because these estimates will likely be different from the estimates for MTBC detection due to the detection of different and multiple gene targets. In this assessment, the testing of the most common resistance mutations (eg, *katG* 315ACC for INH resistance; *rpoB* 531TTG for RIF resistance; *gyrA* 94GGC for FQ resistance; *rplC* T460C for linezolid resistance; *pncA* a-11g for pyrazinamide resistance; and *rrs* A1401G for SLI resistance detection) may be used to establish LoD for resistance testing to some of the first- and second-line compounds detectable by the assay [23]. Line probe assays (LPAs) may be included as comparators for relevant drug compounds. Ideally, the results of this testing should confirm that assay LoD for resistance detection is equivalent or superior to WHO-endorsed comparators.

It should be noted that only genetically and phenotypically well characterized and quantified samples should be used for LoD assessments. In the absence of a WHO international standard, a standardized panel for dynamic range and LoD determination is available from FIND and Zeptometrix [24]. Assay developers should consider using at least 1 drug-sensitive and 1 drug-resistant strain from such a panel for the LoD assessment of DST assays, to increase confidence in the LoD of the assay for both MTBC and resistance detection.

Detection of Resistance-Conferring Mutations

The ability to detect mutations in resistant strains should also be assessed during this analytical study. This is necessary, because no reasonably sized clinical study will be able to achieve a sufficient diversity of strains and resistance-conferring mutations to adequately challenge assay performance. Ideally, assays should be challenged against mutation panels that include high-confidence resistance mutations of notable global prevalence, covering approximately 80%–90% of known resistance mechanisms for any drug. For example, the *katG* 315ACC and *inhA* C-15T mutations represent 80.8% of global INH resistance mechanisms, according to recent mutation grading data [23]. Ideally, 3 independent strains from different WHO regions should be tested for each mutation to guarantee high, reproducible DST assay performance for an epidemiologically diverse set of strains with these relevant resistance mutations. It should be noted that this assessment requires testing against strain panels that have been phenotypically characterized with WHO-endorsed assays, including phenotypic DST on solid or liquid media at the recommended critical concentrations [25], and sequencing to define the genetic basis of resistance. Existing, high-quality WHO strain banks include the FIND TB Strain Bank, with a diversity of genetically and clinically well characterized resistant strains and matched clinical samples [26], as well as the Institute of Tropical Medicine (Antwerp, Belgium) [27, 28], which houses a wide range of MTBC isolates.

Inclusivity and Exclusivity

Analytical studies should also assess DST assay inclusivity and exclusivity, testing assay reactivity against a range of MTBC variants (inclusivity) as well as against other organisms (exclusivity). Although the use of an epidemiologically diverse set of strains during mutation challenge experiments can generate data regarding assay inclusivity, care should be taken to

ensure that the assay has been adequately challenged against MTBC variants. It should also be confirmed that the assay identifies all different MTBC members as TB. For exclusivity testing, a range of NTM and Gram-positive and -negative bacteria, especially those present in oral flora and sputum, should be tested [29]. It is recommended that at least 20 clinically relevant NTMs and at least 10 other bacteria should be tested during this assessment with no observed cross-reactions. Such a panel is available through the European Reference Laboratory Network for TB, and another is currently under development to be available via FIND. This testing may be complemented by an in silico assessment of sequence data, looking at the cross-reactivity of assay primer and probes with all known clinically relevant NTM and pathogens. Finally, interference effects of NTM or human deoxyribonucleic acid (DNA) in mixed samples should be evaluated to confirm assay functionality in these cases [5].

Heteroresistance Detection

Heteroresistance detection should also be assessed in early analytical studies. Replicate mixtures of wild-type and mutant strains or DNA should be tested by the assay within the context of the LoD at set ratios for the most common resistance mutations in each gene region included in the assay [23]. These ratios might be wider or narrower depending on the type of assay tested. For example, a next-generation sequencing technology would likely have a lower threshold for resistance testing, and so a narrower range of mixtures (eg, 0.5%, 1%, 2%, 5%, and 10% mutant:wildtype) may be tested, compared with a real-time assay (eg, 10%, 25%, 50%, 75%, and 90% mutant:wildtype).

CLINICAL EVALUATION STUDIES

After confirming adequate analytical performance, clinical studies should be conducted to evaluate the diagnostic performance of TB DST assays on patient specimens in settings of intended use. The data generated from these studies on assay accuracy can be correlated with early LoD and analytical performance data. Clinical studies will further contribute substantially to ascertainment of diagnostic specificity, support data generated during exclusivity testing, and provide data on operational characteristics to guide policies for use.

Population and Setting

Clinical evaluations should be conducted in diverse settings representative of the TB epidemic in high-burden countries to ensure the study population closely reflects the target population in settings of intended use. Ideally, at least 3 sites from different WHO regions should be selected for these studies. The selection of sites in diverse geographical regions will also ensure that data on operational characteristics are reflective of special issues that may be encountered in different settings. Patient enrollment based upon DR-TB risk factors (eg, previously treated TB patients) can also be an acceptable strategy to enrich for patients with M/XDR-TB in DST studies (additional details on sample size given below).

Reference Standard and Comparators

Given the importance of accuracy estimates in guiding clinical decisions and directing the development of diagnostic algorithms and clinical guidelines, it is imperative that a sound reference standard and informative comparators are incorporated into evaluation studies (Table 4). Currently, culture-based DST methods are the best available reference standard for

MTBC and resistance detection, but these methods are not always reproducible or accurate, particularly for resistance detection [30, 31]. Although genotypic methods such as sequencing may be considered a reliable method to confirm the presence of mutations detectable by TB DST, not all genetic resistance mechanisms are known for every drug and some mutations might not be associated with resistance.

Table 4. Advantages and Disadvantages of Reference Standards and Comparators for DR-TB Diagnostic Evaluation Studies

Test	Type	Advantages	Disadvantages
Phenotypic DST: MGIT960 Löwenstein-Jensen Middlebrook 7H10 Middlebrook 7H11	(WHO-endorsed) Reference Standard	Widely available in most settings May show resistance in a clinical sample where the mutation is unknown or not identified by targeted sequencing Clinical relevance of results is well established	Slow: weeks to months to yield results Difficult to obtain accurate results for certain drugs (eg, PZA) Testing may miss heteroresistance or mutations that confer resistance below the critical concentration (ie, MIC testing would afford better resolution of phenotypic resistance) Biosafety-related issues
Sequencing: Sanger sequencing Targeted next-generation sequencing Whole-genome sequencing Pyrosequencing	Reference Standard	Rapid results compared with culture Detects mutations that may not test resistant at phenotypic DST at a single critical concentration Detection of heteroresistance	Not all resistance-conferring mutations are known May be cost prohibitive in certain settings Need for high-quality DNA for whole-genome sequencing requires a culture step Lack of international standards for NGS quality (eg, minimal coverage requirements) Complicated and nonvalidated NGS analysis pipelines
Phenotypic DST and sequencing	Composite Reference Standard	Combines genetic and phenotypic data to provide a fuller picture of resistance	Slow due to the same need for growth-based results NGS may be cost prohibitive in certain settings
Line probe assays	Comparator	Rapid results Assess the major INH, RIF, FQ, and SLI resistance-conferring gene regions that will be included in all molecular assays	May need to test multiple LPAs (eg, both Hain MTBDR*plus* and MTBDR*sl*) to get full DST comparator information Open assay increases contamination potential Requires manual interpretation Do not detect additional resistance mutations outside of gene target hotspots (eg, the *rpoB* 491 codon mutations)
Xpert MTB/RIF and MTB/RIF Ultra assays	Comparator	Rapid results Assess the major RIF resistance-conferring gene regions that will be included in all molecular assays Automatic interpretation	Limited DST information provided (only RIF resistance)

Abbreviations: DR-TB, drug-resistant tuberculosis; DST, drug-susceptibility testing; FQ, fluoroquinolone; INH, isoniazid; MGIT, mycobacteria growth indicator tube; MIC, minimum inhibitory concentration; NGS, next-generation sequencing; PZA, pyrazinamide; RIF, rifampicin; SLI, second-line injectable; WHO, World Health Organization.

Given the need for a comprehensive picture of drug resistance and the resolution of discrepancies between the index test and reference standard in diagnostic accuracy studies, the use of a composite reference standard, combining genotypic sequencing information and phenotypic DST results, is highly recommended. The benefit of a composite reference standard is that it helps overcome the limitations of individual reference tests: if a specimen is resistant according to phenotypic DST or has a known resistance-conferring mutation, the specimen is classified as drug-resistant, but if both phenotypic DST and sequencing indicate susceptibility, the specimen is classified as drug-susceptible. Because specificity of both phenotypic DST and sequencing is high, this creates a more robust reference standard and allows for a more comprehensive picture of diagnostic assay performance, as seen in the recent evaluation of the Hain MTBDR*plus* Version 2 (Hain Lifescience GmbH, Nehren, Germany) and Nipro NTM+MDRTB (Nipro Corporation, Osaka, Japan) LPAs [32].

The inclusion of WHO-approved tests as comparators in evaluation studies also provides the ability to benchmark and generate stronger evidence for WHO review [33]. In particular, the inclusion of WHO-endorsed LPAs and the Xpert MTB/RIF and/or Ultra assays can benefit both analytical and clinical evaluation studies, because these assays will likely target the

same gene targets as the index tests for first-line and, in the case of the second-line Hain MTBDR*sl* assay (Hain Lifescience GmbH, Nehren, Germany), FQ and SLI resistance detection. Ideally, an evaluation study should assess the diagnostic accuracy of the index test against phenotypic DST, sequencing, and the composite reference standard as well as compare assay diagnostic performance to included comparators.

Sample Size

Sample size is a critical consideration in designing clinical evaluation studies. Sample size should be set to achieve targeted precision for accuracy estimates. Figure 1 shows how the precision of estimates increases as a function of increasing sample size and also demonstrates where increasing precision comes at high cost in terms of the number of patients recruited. For TB DST solutions, these estimates may be based upon TPP performance estimates [5]. Ideally, sensitivity >95% and specificity ≥98% should be achieved compared with sequencing for all drugs included in the assay, in line with minimally acceptable TPP performance characteristics [5]. Furthermore, sensitivity should be >90% for INH, >95% for RIF, and >90% for FQ compared with phenotypic DST, and specificity ≥98% for drug resistance detection for first- and second-line drugs to which the test is able to identify resistance [5], with selected sample size establishing high confidence in obtained diagnostic performance estimates. Ideally, the width of target confidence intervals should be ≤5% for specificity estimates and ≤10% for sensitivity estimates (Table 5). Sample size estimates should also be inflated to account for the number of index and reference test runs expected to yield indeterminate results or errors (eg, <5% according to TPPs). Calculations should also account for the fact that the resistance profiles of enrolled patients will likely vary by site. For example, the anticipated drug resistance profiles of patients enrolled at a DR-TB referral center would vary from those enrolled at a centralized laboratory in a low-TB prevalence setting.

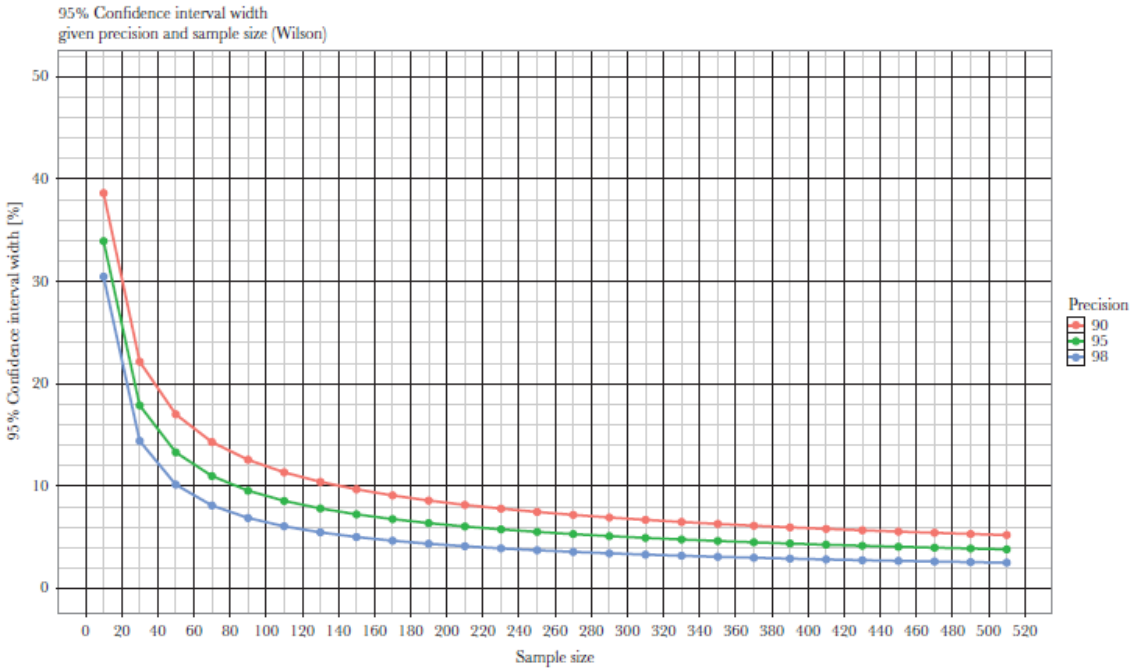

Figure 1. Precision of accuracy estimates as function of sample size. The lines show the precision of accuracy estimates as function of sample size; accuracy point estimates are chosen according to the minimal targets based on the target product profile, ie, sensitivity 98% (blue line) for detecting targeted single-nucleotide

polymorphisms (SNPs) for resistance to rifampicin (RIF), and 95% (green line) for detecting SNPs for resistance to fluoroquinolones (FQs), pyrazinamide (PZA), isoniazid (INH), and aminoglycosides (AGs) and capreomycin when compared with genetic sequencing; specificity 98% (blue line) for any anti-tuberculosis (TB) agent for which the test is able to identify resistance when compared against genetic sequencing; sensitivity 95% (green line) for detecting RIF resistance and 90% (red line) for detecting FQ, PZA, INH, and AG resistance compared with phenotypic cultureThe y-axis shows total width of the 95% confidence interval for sensitivity and specificity for a given sample size. The x-axis shows the necessary number of patients with drug-resistant (DR)-TB to achieve a given precision for sensitivity and the number of patients without DR-TB to achieve a given precision for specificity.

Sample Flow

In designing the clinical evaluation of a TB DST assay, special attention should be paid to sample flow. The index test should be performed on either the raw or processed clinical sample, or both, according to the target starting material for the assay. The reference test and comparator should be performed on the same samples as the index test when possible, to ensure comparability of results. In the case of sputum-based TB DST assays, the index test might be performed twice: once on the direct sample to evaluate assay performance for raw specimens, and once on the cultured isolate along with the reference and comparator methods (Figure 2). This double testing provides estimates of assay performance for 2 different sample types and allows for the resolution of discordant results between tests performed on the raw specimen versus the cultured isolate. Sample flow should also account for minimal sample volume needs for the index, reference, and comparator tests, considering the potential need for repeats of indeterminate results while ensuring that study activities conform with sample collection and processing procedures at clinical sites and do not place undue burden on study participants, as may be the case when acquiring multiple specimens (see more detailed discussion of issues relating to this topic in Paper 2).

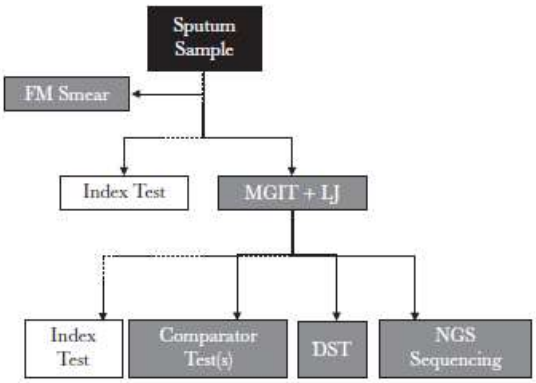

Figure 2. Example of a recommended sample flow for clinical evaluation studies. The graphic reflects sample flow for the evaluation of a tuberculosis (TB) drug-susceptibility test (DST) solution. All molecular assays should be performed in accordance with standard operating protocols provided by the manufacturers, whereas specimen handling and processing should be carried out according to National TB Programme policies and standards. After screening of patients or samples for inclusion in the study, sputum samples should be collected and acid-fast bacilli smear and mycobacteria growth indicator tube (MGIT) and Löwenstein-Jensen (LJ) culture should be performed directly for all samples in addition to the index test. After direct testing, phenotypic MGIT DST should be performed for all culture-positive samples for relevant drug compounds. Cultured samples will also undergo subsequent molecular testing by comparator assays and targeted next-generation sequencing (NGS) of relevant gene regions and another index test. The NGS should be performed from the culture isolate to obtain sequencing reads of high quality. FM, fluorescent microscopy.

Key Issues Beyond Accuracy

There is the additional need to monitor and evaluate other aspects of assay performance that must be planned for when designing clinical studies. Data on assay operation should be collected through observed usage and user appraisal questionnaires during clinical trials to ultimately guide policies regarding assay use. Various assay technical and operational performance parameters that should be measured during clinical studies include time-to-result, indeterminate rates, and other factors, which may also be assessed outside a clinical study, are listed in Table 2. The need for sample referral and data transmission networks should also be noted, particularly for centralized DST solutions. In addition, operators should take special care in monitoring and noting potential cross-contamination issues when incorporating the DST assays into laboratory flow. Finally, there are important considerations for interpreting reference standard and comparator results from a clinical study. Notably, the potential for the different technologies to detect heteroresistance in clinical samples may be a concern, especially for drugs for which resistant subpopulations are commonly observed, such as the FQs [34, 35], where the frequency of the resistant allele may be below the threshold of detection of different diagnostic assays. These populations may also be grown out either resistant or susceptible by phenotypic DST, which could lead to discordances between phenotypic DST and sequencing if sequencing is only performed on the culture isolate. For this reason, researchers may consider conserving sputum samples to later perform targeted deep sequencing for discordance resolution. This ensures that the most accurate genotypic reference data is obtained for clinical samples, especially when the index test is being performed directly on the clinical sample.

For the analysis of MTBC detection, it is important that results are analyzed by smear status and by patient TB history to account for possible false-positive results due to remnant TB in samples and to ensure that clinical data reflect accurate MTBC detection estimates. These considerations are vital to ensure accurate reporting of study results, including key assay performance parameters (detailed discussion of issues relating to this topic is given in Paper 2).

CONCLUSIONS

As novel tests are developed in line with existing TPPs, appropriate evidence generated on their performance and operational characteristics should rapidly follow. This will ensure that policy recommendations can be made, which is the first step to access these tests for TB patients. Studies following the outline provided here are expected to generate high-quality laboratory and clinical data regarding assay performance and operational characteristics and support WHO review and potential recommendation.

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.

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.

Potential conflicts of interest. All authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

References

1. World Health Organization. *Global Tuberculosis Report 2018*. Geneva: World Health Organization; 2018.
2. Shah NS, Richardson J, Moodley P, et al. Increasing drug resistance in extensively drug-resistant tuberculosis, South Africa. *Emerg Infect Dis* 2011; 17:510–513.
3. Kendall EA, Azman AS, Cobelens FG, Dowdy DW. MDR-TB treatment as prevention: the projected population-level impact of expanded treatment for multidrug-resistant tuberculosis. *PLoS One* 2017; 12:e0172748.
4. Sharma A, Hill A, Kurbatova E, et al. Estimating the future burden of multidrug-resistant and extensively drug-resistant tuberculosis in India, the Philippines, Russia, and South Africa: a mathematical modelling study. *Lancet Infect Dis* 2017; 17:707–715.
5. World Health Organization. *Meeting Report: High-Priority Target Product Profiles for New Tuberculosis Diagnostics: Report of a Consensus Meeting*. Geneva: World Health Organization; 2014.
6. Gilpin C, Korobitsyn A, Weyer K. Current tools available for the diagnosis of drug-resistant tuberculosis. *Ther Adv Infect Dis* 2016; 3:145–151.
7. Steingart K, Schiller I, Horne DJ, Pai M, Boehme C, Dendukuri N. Xpert MTB/RIF assay for pulmonary tuberculosis and rifampin resistance in adults. *Cochrane Database Syst Rev* 2014; CD009593.
8. Scott L, David A, Noble L, et al. Performance of the Abbott Realtime MTB and MTB RIF/INH assays in a setting of high tuberculosis and HIV coinfection in South Africa. *J Clin Microbiol* 2017; 55:2491–2501.
9. Park JE, Huh HJ, Koh WJ, Song DJ, Ki CS, Lee NY. Performance evaluation of the Cobas TaqMan MTB assay on respiratory specimens according to clinical application. *Int J Infect Dis* 2017; 64:42–46.
10. Hillemann D, Haasis C, Andres S, Behn T, Kranzer K. Validation of the FluoroType® MTBDR assay for the detection of rifampicin and isoniazid resistance in *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol* 2018; 56:pii:e00072-18.
11. Johns Hopkins Center for Clinical Global Health Education. *Multicenter study of the accuracy of the BD MAX MDR-TB assay for detection of M. tuberculosis complex and mutations associated with resistance to rifampin or isoniazid*. 2017. Available at: https://main.ccghe.net/content/mul-ticenter-study-accuracy-bd-max-mdr-tb-assay-detection-m-tuberculosis-complex-and. Accessed 26 February 2019.
12. Xie YL, Chakravorty S, Armstrong DT, et al. Evaluation of a rapid molecular drug-susceptibility test for tuberculosis. *N Engl J Med* 2017; 377:1043–1054.
13. *FIND. Diagnostic Pipeline Tracker*. Available at: https://www.finddx.org/tb/pipeline/. Accessed 26 February 2019.
14. Hofmann-Thiel S, Molodtsov N, Antonenka U, Hoffmann H. Evaluation of the Abbott Realtime MTB and realtime MTB INH/RIF assays for direct detection of *Mycobacterium tuberculosis* complex and resistance markers in respiratory and extrapulmonary specimens. *J Clin Microbiol* 2016; 54:3022–3027.
15. Roche Diagnostics. *Leading the way in Integrated Diagnostic Solutions*. Available at: https://www.roche.com/dam/jcr:c4a05bd6-8567-4c3e-90c1-eb6b564b1780/en/irp20180801.pdf. Accessed 26 February 2019.
16. Rice JE, Reis AH Jr, Rice LM, Carver-Brown RK, Wangh LJ. Fluorescent signatures for variable DNA sequences. *Nucleic Acids Res* 2012; 40:e164.

17. Dolinger DL, Colman RE, Engelthaler DM, Rodwell TC. Next-generation sequencing-based user-friendly platforms for drug-resistant tuberculosis diagnosis: a promise for the near future. *Int J Mycobacteriol* 2016; 5(Suppl 1):27–28.
18. Nikam C, Jagannath M, Narayanan MM, et al. Rapid diagnosis of *Mycobacterium tuberculosis* with Truenat MTB: a near-care approach. *PLoS One* 2013; 8:e51121.
19. *FIND. DX Pipeline*. Available at: https://www.finddx.org/dx-pipeline-status/. Accessed 26 February 2019.
20. World Health Organization. *The Use of Next-Generation Sequencing Technologies for the Detection of Mutations Associated with Drug Resistance in Mycobacterium tuberculosis Complex: Technical Guide*. Geneva: World Health Organization; 2018.
21. World Health Organization. *WHO Treatment Guidelines for Multidrug- and Rifampicin-Resistant Tuberculosis. 2018 Update*. Geneva: World Health Organization; 2018.
22. World Health Organization. *WHO Recommends New Tuberculosis Test*. Geneva: World Health Organization; 2016.
23. Miotto P, Tessema B, Tagliani E et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis. Eur Respir J* 2017; 50:1–13.
24. *FIND. Samples and reference materials*. Available at: https://www.finddx.org/specimen-banks/. Accessed 26 February 2019.
25. World Health Organization. *Technical Report on Critical Concentrations for TB Drug Susceptibility Testing of Medicines Used in the Treatment of Drug-Resistant TB*. Geneva: World Health Organization; 2018.
26. Tessema B, Nabeta P, Valli E, et al. FIND tuberculosis strain bank: a resource for researchers and developers working on tests to detect *Mycobacterium tuberculosis* and related drug resistance. *J Clin Microbiol* 2017; 55:1066–1073.
27. Vincent V, Rigouts L, Nduwamahoro E, et al. The TDR Tuberculosis Strain Bank: a resource for basic science, tool development and diagnostic services. *Int J Tuberc Lung Dis* 2012; 16:24–31.
28. Nathanson CM, Cuevas LE, Cunningham J, et al. The TDR Tuberculosis Specimen Bank: a resource for diagnostic test developers. *Int J Tuberc Lung Dis* 2010; 14:1461–1467
29. U.S. Department of Health and Human Services. *Class II special controls guideline: Nucleic acid-based in vitro diagnostic devices for the detection of Mycobacterium tuberculosis complex and genetic mutations associated with Mycobacterium tuberculosis complex antibiotic resistance in respiratory specimens.* 2014 Available at: https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM419468.pdf. Accessed 26 February 2019.
30. Kim SJ. Drug-susceptibility testing in tuberculosis: methods and reliability of results. *Eur Respir J* 2005; 25:564–569.
31. Van Deun A, Aung KJ, Bola V, et al. Rifampin drug resistance tests for tuberculosis: challenging the gold standard. *J Clin Microbiol* 2013; 51:2633–2640.
32. Nathavitharana RR, Hillemann D, Schumacher SG, et al. Multicenter noninferiority evaluation of hain genotype MTBDRplus version 2 and Nipro NTM+MDRTB line probe assays for detection of rifampin and isoniazid resistance. *J Clin Microbiol* 2016; 54:1624–1630.
33. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013; 158:544–554.
34. Kaplan G, Post FA, Moreira AL, et al. *Mycobacterium tuberculosis* growth at the cavity surface: a microenvironment with failed immunity. *Infect Immun* 2003; 71:7099–7108.
35. Zhang X, Zhao B, Liu L, Zhu Y, Zhao Y, Jin Q. Subpopulation analysis of heteroresistance to fluoroquinolone in *Mycobacterium tuberculosis* isolates from Beijing, China. *J Clin Microbiol* 2012; 50:1471–1474.
36. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927; 22L:209–212