The Journal of Infectious Diseases

SUPPLEMENT ARTICLE

Guidance for Studies Evaluating the Accuracy of Sputum-Based Tests to Diagnose Tuberculosis

Samuel G. Schumacher,[1] William A. Wells,[2] Mark P. Nicol,[3] Karen R. Steingart,[4] Grant Theron,[5] Susan E. Dorman,[6] Madhukar Pai,[7] Gavin Churchyard,[8,9,10] Lesley Scott,[11] Wendy Stevens,[11] Pamela Nabeta,[1] David Alland,[12] Karin Weyer,[13] Claudia M. Denkinger,[1,14] and Christopher Gilpin[13]

[1]FIND, Geneva, Switzerland; [2]United States Agency for International Development, Washington, District of Columbia; [3]School of Biomedical Sciences, University of Western Australia, Perth, Australia [4]Liverpool School of Tropical Medicine, United Kingdom; [5]DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, SA MRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa; [6]Medical University of South Carolina, Charleston; [7]McGill International TB Centre and Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada; [8]Aurum Institute, Cape Town, South Africa; [9]School of Public Health, University of the Witwatersrand, Johannesburg, South Africa; [10]Advancing Care and Treatment for TB/HIV, South African Medical Research Council, Parktown, South Africa; [11]University of the Witwatersrand and National Health Laboratory Service, Johannesburg, South Africa; [12]Rutgers University, New Jersey; [13]World Health Organization, Geneva, Switzerland; [14]University Hospital Heidelberg, Division of Tropical Medicine, Centre of Infectious Diseases, Germany

Tests that can replace sputum smear microscopy have been identified as a top priority diagnostic need for tuberculosis by the World Health Organization. High-quality evidence on diagnostic accuracy for tests that may meet this need is an essential requirement to inform decisions about policy and scale-up. However, test accuracy studies are often of low and inconsistent quality and poorly reported, leading to uncertainty about true test performance. Here we provide guidance for the design of diagnostic test accuracy studies of sputum smear-replacement tests. Such studies should have a cross-sectional or cohort design, enrolling either a consecutive series or a random sample of patients who require evaluation for tuberculosis. Adults with respiratory symptoms are the target population. The reference standard should at a minimum be a single, automated, liquid culture, but additional cultures, follow-up, clinical case definition, and specific measures to understand discordant results should also be included. Inclusion of smear microscopy and Xpert MTB/RIF (or MTB/RIF Ultra) as comparators is critical to allow broader comparability and generalizability of results, because disease spectrum can vary between studies and affects relative test performance. Given the complex nature of sputum (the primary specimen type used for pulmonary TB), careful design and reporting of the specimen flow is essential. Test characteristics other than accuracy (such as feasibility, implementation considerations, and data on impact on patient, population and health systems outcomes) are also important aspects.

Keywords. diagnostics; study design guidance; target product profiles; tuberculosis; WHO End TB strategy.

For decades, sputum smear microscopy has been used as the initial test to detect active tuberculosis (TB). Despite its ubiquity, microscopy is suboptimal because it has low sensitivity, high interoperator variability, is largely unhelpful in extrapulmonary and childhood TB, and does not detect drug resistance [1–7].

Xpert MTB/RIF (Xpert; Cepheid, Sunnyvale, CA) overcame some shortcomings of sputum smear microscopy, based on its increased sensitivity and simultaneous detection of resistance to rifampicin [8]. Xpert is recommended by the World Health Organization (WHO) to be used as the initial test rather than microscopy for all persons with signs and symptoms of TB [9], and there has been substantial uptake by high burden countries [10].

However, 2 major barriers remain. First, Xpert is primarily suited for placement at the district hospital level or higher, which are above the subdistrict location where most smear microscopy is performed [11]. There remains no single rapid, accurate, and robust TB diagnostic test suitable for use at the subdistrict location. Second, in all but a very few high TB burden countries, the high cost of the instruments and maintenance and the costs of cartridges have prevented full adoption of this test (ie, its use as an initial test for all patients presenting with signs and symptoms of active TB) [12].

In an ideal setting, sputum would be replaced by a specimen that is easier to collect with less variability in quality (considerations for developing biomarker-based assays using nonsputum specimens are outlined further in the paper by Drain et al in this series [13]). However, sputum is likely to remain a crucial specimen for the immediate future because (1) currently no accurate nonsputum biomarker-based tests are available for TB, and (2) even if accurate nonsputum biomarker-based tests become available, drug-resistance testing will likely remain a necessity but may not be feasible with tests that are not based on pathogen deoxyribonucleic acid (DNA). Decentralized testing for TB also remains a priority in many settings because most TB patients present at primary care centers, specimen transport is challenging, and pretreatment loss to follow-up is common if there are diagnostic delays [14]. Thus, the development of a

Correspondence: S. G. Schumacher, PhD, FIND, Campus Biotech, Chemin des Mines 9, 1202 Geneva, P.O. Box 87, 1211 Geneva 20, Switzerland (samuel.schumacher@finddx.org).

The Journal of Infectious Diseases® 2019;220(S3):S99–107

DOI: 10.1093/infdis/jiz258

rapid, accurate, and simple smear-replacement test that can be implemented where patients first present for diagnosis remains a high priority [15]. Such a test should facilitate the initiation of appropriate treatment during the same clinical encounter or the same day.

In 2014, the WHO and partners developed target product profiles (TPPs) for new TB diagnostics, describing the minimal and optimal performance and operational characteristics of tests for high-priority needs, including a smear-replacement test that could be used in microscopy centers [15]. Microscopy centers are defined here as primary healthcare centers with attached peripheral laboratories with minimal infrastructure, and they are typically present at the subdistrict level, although microscopy may be done throughout a tiered network. At a minimum, a suitable test should (1) have high (>98%) specificity so that positive results can be used to rule-in TB, (2) have a higher sensitivity than sputum smear microscopy (>60% for smear-negative TB) to enable earlier detection of pulmonary TB, (3) be robust enough for use in microscopy centers (or comparable basic healthcare facilities) under challenging environmental conditions (temperature, humidity, dust, limited infrastructure), (4) be simple enough to be performed by healthcare workers with minimum training, and (5) have low cost (<$6 per test) to enable large-scale use. In an ideal setting, such a test should (1) have even higher (>95%) sensitivity, (2) cost less (<$4 per test), and (3) permit the monitoring of treatment response and drug-susceptibility testing (DST) [14] (see detailed discussion of this topic in the paper by Georghiou et al in this series [16]). Several emerging technologies (Table 1 [17]) have the potential to meet the need of a smear-replacement test, but developing a simple, affordable instrument that can meet the needs in microscopy centers remains challenging [18].

The goal of this article is to provide guidance for studies evaluating the diagnostic accuracy of sputum-based tests to diagnose TB. Although the main focus is on the evaluation of decentralized technologies, many considerations apply equally to more centralized testing systems. We summarize our recommendations in Table 2.

GENERAL STUDY DESIGN CONSIDERATIONS

To obtain unbiased and precise estimates of sensitivity and specificity, clinical studies evaluating diagnostic test accuracy should use a cross-sectional or cohort study design, enrolling a sufficient number of consecutive or randomly selected patients requiring TB evaluation (Figure 1). However, before undertaking resource-intensive prospective evaluations, case-control studies using banked specimens from well characterized cohorts and/or studies involving negative sputum specimens spiked with known numbers of *Mycobacterium tuberculosis* (MTB) bacilli may be performed first. It is important to note that case-control studies should avoid comparing severe cases to healthy controls that can result in overestimations of test accuracy (spectrum bias). Although such "proof-of-concept" studies are not a major focus of this document, investigators should be aware that these types of studies can play an important role in the early assessment of smear-replacement tests, particularly if they include head-to-head studies against assays with well established performance characteristics. If banked specimens are processed and stored appropriately, these specimens can be used to evaluate DNA-based tests. Once promising smear-replacement tests have been identified, they should be evaluated in clinical studies using fresh specimens collected and processed under routine conditions.

POPULATION AND SETTING

The target population for initial accuracy studies of a new smear-replacement test should be adults self-presenting with respiratory symptoms suggestive of TB (ie, passive case finding), including people living with HIV (PLHIV). For patients without HIV, cough ≥2 weeks is used to identify patients with suspected TB [19, 20], whereas less stringent criteria (cough of any duration, fever, night sweats, or weight loss) is used for PWH and other high-risk groups [21]. Adults with suspected pulmonary TB represent the optimal initial study population because (1) the reference standard (culture) has good sensitivity in this patient group, (2) it represents the largest proportion of the target population to which the test would later be applied in practice,

Table 1. Technologies That Have the Potential to Meet the Need of a Smear-Replacement Test as Defined in the WHO TPP [17]

Assay-Instrument Combinations Commercially Available in 2018	Assay-Instrument Combinations Expected to Launch in 2019	Companies Developing Assay-Instrument Combinations With Potential to Meet the TPP[a]
• Molbio's Truenat MTB assay used with the Trueprep DNA extraction device and Truelab PCR analyser[b] • GeneXpert Edge used with Xpert MTB/RIF or Xpert MTB/RIF Ultra[d]	• Cepheid GeneXpert Omni used with Xpert MTB/RIF or Xpert MTB/RIF Ultra[c]	• Ustar Biotechnologies • QuantuMDx • Bioneer • Akonni • SelfDiagnostics

Abbreviations: DNA, deoxyribonucleic acid; MTB, *Mycobacterium tuberculosis*; PCR, polymerase chain reaction; TPP, target product profile; WHO, World Health Organization.

[a]These are still not close to commercialization. Note that Alere Q was a promising development that has been discontinued.

[b]For the Molbio system, in its current form precision pipetting is needed, a separate DNA extraction and DNA amplification/detection device pose cross-contamination risks, and the data available on its accuracy are limited.

[c]Omni will run the Xpert MTB/RIF and Xpert MTB/RIF Ultra assays, which have good diagnostic accuracy, and are recommended by the WHO. No data are available yet.

[d]For the GeneXpert Edge system, a dust filter and a battery will allow broader use compared with the other GeneXpert systems, while the limitations for use at high temperatures will remain.

Table 2. Overview of Recommendations for Study Design

Topic	Recommendation
General study design	• Use a cross-sectional or cohort study enrolling either a consecutive series or a random sample of patients who require evaluation for TB (avoid using known, severe cases, and healthy controls, because this introduces spectrum bias and can lead to overestimates of test accuracy) • Aim at a sample size that ensures at least 60 patients with smear-negative, culture-positive TB are included; smaller studies are still valuable and can be integrated in a systematic review and meta-analysis • Follow STARD as well as the more detailed advice contained in this guidance for reporting
Population and setting	• Avoid selecting patients in whom TB has already been diagnosed by another test or who have already started on TB treatment • For initial studies focus on adults, including PWH, who have respiratory symptoms suggestive of TB; subsequently evaluate other key groups (eg, children, extrapulmonary TB, patients identified through active case finding) • Ideally recruit patients at primary healthcare centers • Report TB prevalence and proportion of smear-negative, culture-positive TB (among all culture-positive TB) for each patient recruitment site • Perform testing in highly proficient laboratories in initial studies; testing in intended use setting should only be done if testing quality can be guaranteed • Provide stratified accuracy estimates for key subpopulations (by HIV status and smear status)
Index test	• Consider specifics of the index test under investigation: • For tests with nonautomated readout, blinding is essential to make sure the index test is interpreted independently of the reference test or comparators • For tests that incorporate testing for drug-resistance, pay attention to additional considerations [16]
Reference standard and comparators	• Use automated liquid culture as the reference standard, optimally more than 1 culture done from specimens taken on separate days • Avoid partial or differential verification bias, ie, all those who received the index test should also receive the same reference standard • Include follow-up, clinical case definition, and additional measures to understand discordant (index-test-positive, culture-negative) results • Include smear microscopy and Xpert MTB/RIF (or MTB/RIF Ultra) as comparators
Flow and specimen issues	• Carefully design and report the study sample flow, considering the limitations of each approach (see Table 3) • In many cases, performing the index test, comparator and reference standard from a homogenized native sputum specimen is the preferred option for the specimen flow
Key issues beyond accuracy	• Test characteristics other than diagnostic accuracy are also critical and need to be assessed systematically as well • Implementation studies can help identify bottlenecks that need to be overcome if improved accuracy of new tests is to be capitalized upon • The potential clinical and population level impact of new tests needs to be assessed through modeling and empirical studies

Abbreviations: HIV, human immunodeficiency virus; PLHIV, people living with HIV; TB, tuberculosis. index test, test under investigation

and (3) sufficient volume of sputum can usually be obtained from such patients. Patients in whom TB has already been diagnosed by another test or who have already started on TB treatment should be excluded, because enriching with patients that are positive by sputum smear microscopy or Xpert will lead to overestimates of sensitivity of the test under investigation.

Children and patients with extrapulmonary and early-stage TB are other important patient groups in whom accuracy needs to be determined, typically in separate and subsequent studies. Because they may have low numbers of bacilli in respiratory secretions or other specimens, sensitivity of a test is commonly lower than that obtained when testing sputum from adults with respiratory symptoms. Early-stage TB may also be encountered, for example, due to early presentation of patients to diagnostic clinical services or because patients were identified during screening or active case finding. For example, one important use for this will likely be the case when a smear-replacement test is used as the confirmatory test for those who are asymptomatic but screen positive by chest x-ray in an outreach setting.

Sensitivity of sputum-based tests depends on the bacillary burden of MTB in sputum specimens, and therefore presenting sensitivity estimates separately by smear status is essential to gauge performance in the most difficult-to-diagnose patients and to estimate the potential incremental yield over conventional sputum smear microscopy [22]. Providing accuracy estimates for PWH, children, or patients with early disease separately is also important (even if numbers are small), to allow inclusion in meta-analyses. Studies focusing specifically on these patient groups are also needed as a next step, once performance in adults with respiratory symptoms has been established.

In addition to the case-finding strategy (passive vs active case finding), test sensitivity is also influenced by the recruitment setting (community, clinic, hospital), which reflects the spectrum of TB disease severity in a population; pauci-bacillary TB will be more prevalent among patients undergoing clinic- or community-based case finding, relative to patients requiring hospitalization or self-presenting to clinics for respiratory symptoms. In an ideal setting, initial studies of novel smear-replacement tests will recruit patients self-presenting to primary healthcare centers with TB symptoms, to help ensure that the patient spectrum reflects both the case-finding strategy and clinical setting of intended test use. Pulmonary TB patients diagnosed in the outpatient setting, and especially during active case finding, tend to be relatively early in their disease process and thus have low bacillary burdens, a scenario that tends to drive down investigational test sensitivity compared with culture but

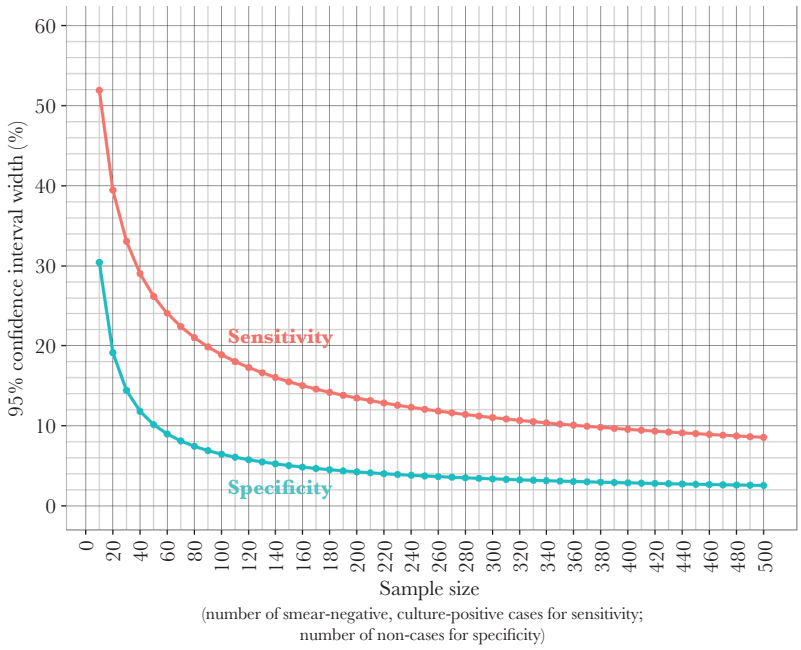

Figure 1. Precision of accuracy estimates as function of sample size. The lines show the precision of accuracy estimates as a function of sample size, when sensitivity (blue line) and specificity (red line) are fixed at the minimum targets (60% sensitivity among smear-negatives, 98% specificity) established by the target product profile (TPP). The y-axis shows total width of the 95% confidence interval (CI) (ie, upper limit of the 95% CI minus the lower limit of the 95% CI) for sensitivity and specificity for a given sample size. The x-axis shows the number of smear-negative tuberculosis (TB) cases and non-TB cases needed to achieve a given precision for sensitivity and specificity, respectively. Sensitivity among smear-negative TB patients is shown here, rather than overall sensitivity, because (1) sensitivity for detecting this group is a crucial performance target in the TPP and (2) this group represents a small subset of all patients enrolled and thus drives sample size needs. Studies of novel smear-replacement tests should aim to enroll ≥60 smear-negative, culture-positive TB patients [23]. Assuming 30% smear-negative, culture-positive TB prevalence, 200 culture-positive TB cases (assuming no losses or exclusions) would be required to obtain a sensitivity estimate with a 24% 95% CI width. This figure also shows that increasing the sample size beyond 60 smear-negative culture-positive TB cases, yields diminishing returns in terms of narrowing the CI width.

augment incremental yield over sputum smear microscopy [24]. In addition, lower TB prevalence in the outpatient setting means that high assay specificity is critical to ensure that test positive predictive value remains sufficiently high. To facilitate assessment of patient spectrum, we recommend reporting, for each patient recruitment site, TB prevalence and proportion of TB cases that were smear-negative, culture-positive [22].

The site of patient recruitment may be different from the site where testing is conducted, but close attention must then be paid to appropriate sample transport to avoid high contamination rates. Initial data are usually best generated via testing in controlled laboratory settings, eg, in a few reference laboratories, to assess diagnostic accuracy under "ideal" conditions in a controlled environment (temperature, humidity, dust), stronger infrastructure (electricity, connectivity), and experienced staff (eg, with prior training on use of molecular methods and good laboratory practices to prevent cross-contamination events). Reference laboratories also typically allow easier access to optimal reference standard testing, facilities for resolution of discordant results, and the ability to test a large number of specimens in a standardized manner. While data from testing in settings of intended use are also critical to ensure consistent performance under more challenging conditions, this might only be possible in later implementation studies.

INDEX TEST

Clear reporting of how the index test (the test under investigation) is performed is essential, as is clear reporting on indeterminate and invalid results or instrument failures. Certain considerations may be important depending on the specifics of the index test. If a test can process a large sputum input volume, it may be important to allow 1 complete specimen to be tested by that test (ie, no "sharing of that specimen" with other technologies), because this may enable high sensitivity that would not be captured otherwise. A test that incorporates simultaneous DST also requires additional considerations (eg, low limit of detection for DST resistance targets) as discussed in "Paper 5" in this series [16]. If the assay readout is not automated and requires a degree of subjective interpretation, prespecification of cutoffs for positivity and blinding of readers to other test results are essential, and interreader reliability needs to be assessed.

REFERENCE STANDARD AND COMPARATORS

We recommend using at least 1 automated liquid culture as the primary reference standard for diagnostic accuracy studies of smear-replacement tests (please refer to discussion on this in "Paper 1" of this series [25]), and all those who received the index test should also receive the same reference standard to

avoid partial or differential verification bias. It is important to acknowledge (1) that there can be large variability of bacillary load between specimens and even within specimens and (2) that even culture is not a perfect reference standard and thus that, because new assays are becoming increasingly more sensitive, false-negative culture results need to be considered—in particular after lengthy specimen transport or overly harsh decontamination of specimens.

Steps that can be taken to reduce the risk of bias due to limitations of the reference standard are as follows: (1) rigorous implementation of the reference standard, including quality control and quality assurance; (2) aiming for liquid culture contamination rates between 8% and 15% and monitoring these during the study; and (3) using more than a single culture per patient (ideally from multiple specimens obtained on different days) to define the reference standard. A clinical or composite reference standard may also be considered to supplement analyses based on culture, and this is particularly pertinent for pediatric and extrapulmonary TB (see further discussion on this topic in the paper in this series by Drain et al [13]).

Steps that can be taken to understand discordant (index-test-positive, culture-negative) results include the following: (1) thorough in silico analyses and exclusivity studies before study initiation; (2) following-up patients to uncover subsequent culture conversion and examination of alternative diagnoses; (3) environmental testing during the study to assess potential for cross-contamination; (4) sequencing of amplicons to detect potential nonspecific amplification; (5) rigorous assessment of prior treatment for TB; and (6) exploration of other patient- and setting-specific characteristics that may lead to false-positive results. Please see more detailed elaborations of these concepts in the accompanying glossary.

Accuracy estimates will vary between studies not only due to variation in patient spectrum but also as procedures for culture and microscopy vary [26]. For example, sensitivity estimates for the index test will decrease when using liquid rather than solid culture, with increasing number of cultures done, increasing number of specimens on which culture is performed, and increasing number of days on which specimens are obtained. In addition, estimates of sensitivity of the new test among smear-negative patients will be lower when (1) using a more sensitive process for smear analysis (ie, using fluorescence microscopy instead of Ziehl-Neelsen), (2) using multiple smears to classify a patient as smear negative (instead of a single smear), and (3) highly proficient operators are preparing and reading the smears.

In a diagnostic test accuracy study, the reference standard is not a comparator but the method used to determine true disease status, which allows measuring the accuracy of the index test (and accuracy of comparators). Smear microscopy or Xpert or other approved and well studied tests can be utilised as comparator tests. The sensitivity of sputum smear microscopy and Xpert observed in a given study provides a good indication of the studies' patient spectrum. Inclusion of a comparator test also allows for an evaluation of the incremental yield and stratification of sensitivity by the comparator test. Having comparative data on Xpert is extremely useful given the large amount of data available on its diagnostic accuracy. Showing similar or better sensitivity than Xpert, even on a relatively small number of patients, is stronger evidence for good performance than a larger study without this comparator. Comparing the accuracy of multiple index tests that were evaluated in different studies is usually problematic because of variation of the patient spectrum unless the varying patient spectrum between studies can be understood through testing with a comparator test such as Xpert (as discussed in section on "Population and Setting") [27].

FLOW AND SPECIMEN ISSUES

For sputum, the fact that there is important variability (day-to-day, specimen-to-specimen, within-specimen) needs to be taken into consideration when designing the specimen flow of a study. It is important to include a sample flow diagram as part of reporting (see Figure 2 as an example). Testing with the index test on one specimen and comparator test on another specimen (possibly from another day) can make interpretation of results difficult given the sample-to-sample variability. At best this will result in increased random error ("noise") or at worst in bias if the difference between the specimens is systematic. Performing the index test, the comparator, and 1 culture on the same specimen facilitates interpretation of results and can provide the most direct evidence on comparative accuracy, but the large specimen volume requirement and difficulty encountered in splitting viscous samples can make this approach impractical. The sensitivity of a test is partly dependent on the number of bacilli per specimen volume so sputum input volume is also an important parameter. Thus, comparing sensitivity of one test on a native sputum specimen to the sensitivity of another test on a concentrated pellet from a higher input volume is rarely appropriate.

With regard to sputum specimen flow, there are at least 4 different approaches that can be used to achieve the goal of performing more than 1 test (ie, index test, reference test, and in some instances a comparator test) (see Table 3). One option is to apply the index test and comparator test on 2 separate native sputum specimens, allocated through a randomization scheme. This approach completely retains the challenging sputum matrix and thus applicability of data with regards to the intended use. However, this approach requires a very large sample size to yield precise comparative results to account for the potentially large sample-to-sample variability described above. This approach also requires at least 3 specimens (one each for the index test, comparator, and reference standard) and thus usually 2 patient visits.

Alternatively, a participant's sputum specimen can be split physically into 2 or more portions for testing [28, 29]. This will often only be possible in a laboratory. The "splitting" procedure used should be carefully considered, and the methods should

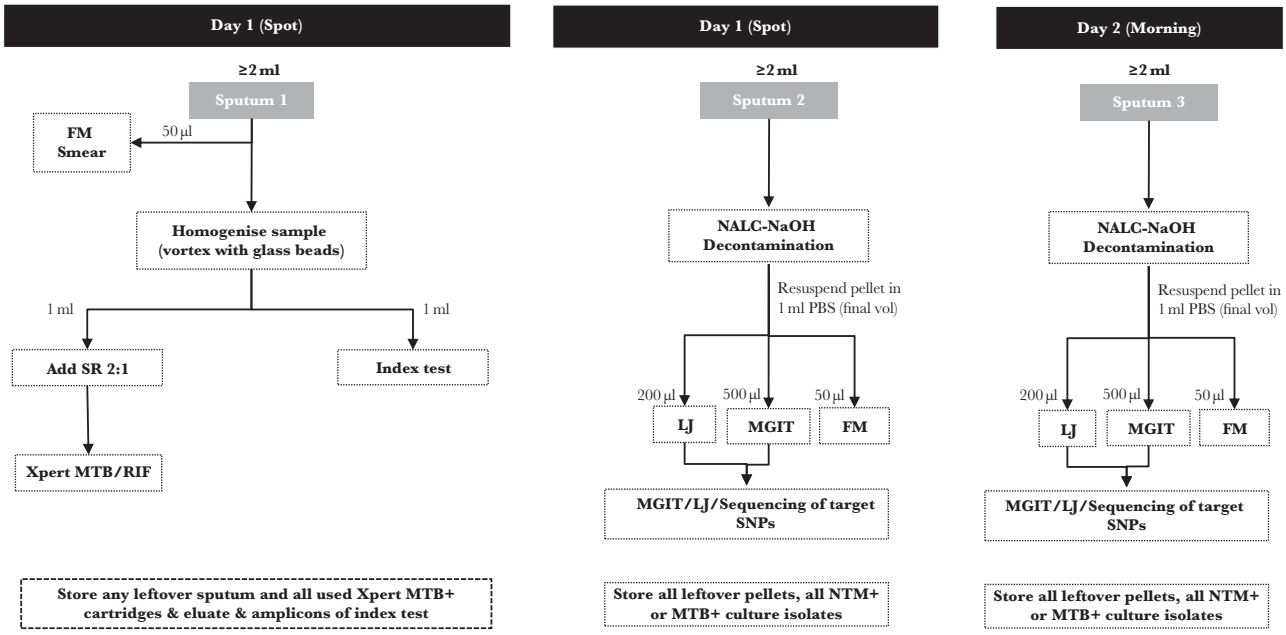

Figure 2. Example of a sample flow diagram for diagnostic accuracy studies a smear-replacement tests. This figure shows an example sputum specimen flow diagram. Studies evaluating the diagnostic accuracy of novel sputum-based tuberculosis (TB) diagnostics should include sputum specimen flow diagrams in their reporting to allow readers to contextualize accuracy estimates. Flow diagrams should include when sputum was collected (spot vs morning), sputum processing methods, and the type and number of TB tests performed from a single specimen. In this hypothetical study, 3 sputum specimens were collected from all patients (2 spot specimens on day 1 and 1 morning specimen on day 2). Sputum 1 underwent fluorescence microscopy (FM) smear before undergoing glass bead homogenization. The homogenized sample was then split for Xpert MTB/RIF testing and the index test. Sputum 2 and 3 undergo identical processing methods and TB testing (FM smear, solid culture, liquid culture); *Mycobacterium tuberculosis* (MTB) culture isolates are then sequenced for target single nucleotide polymorphisms (SNPs). LJ, Lowenstein-Jensen; MGIT, Mycobacterial Growth Indicator Tube liquid culture; NTM, nontuberculous mycobacteria; PBS, phosphate-buffered saline; SR, Xpert sample reagent.

be described in detail. Three options are as follows: (1) split an unhomogenized native sputum specimen and allocate aliquots randomly to different assays; (2) split a homogenized native sputum specimen and allocate aliquots to different assays; or (3) split a concentrated, decontaminated specimen. Testing a native sputum specimen is in line with the intended use of a smear-replacement test on an unprocessed specimen. However, due to within-specimen heterogeneity, the number of bacilli may differ substantially between different aliquots derived from a single specimen, and a larger sample size would be required to compensate for the resulting increase in random error (similar to testing 2 separate specimens). The high viscosity of sputum may make it necessary to homogenize specimens (eg, by vortexing with glass beads) to facilitate physical splitting, and it has the additional advantage of rendering aliquots more homogenous and reducing random variability [30]. On the other hand, homogenization affects the matrix and thus potentially affects assay performance characteristics, and it can introduce contamination.

Table 3. Options for Performing Index and Comparator Test on One or Multiple Specimens

Options	Applicability of Data With Regards to Intended Use (ie, Data Addresses the Challenge of Sputum Matrix)	Risk of Random Error, Difficulty in Interpreting Discordant Results and Sample Size	Comments
Test separate, unhomogenized raw sputum specimens with index test and comparator	High	High	Requires 3 sputum specimens (reference standard, index test, comparator) and thus likely 2 patient visits
Split unhomogenized raw sputum and allocate aliquots at random for testing with index test and comparator	Medium high	Medium high	Splitting unhomogenized raw sputum is practically challenging with viscous samples and limited volumes
Split homogenized raw sputum and test aliquots with index test and comparator	Medium low	Medium low	Great care must be taken to avoid cross-contamination
Split concentrated, decontaminated sputum and test aliquots with index test and comparator	Low	Low	Essential to have evidence to show that index test also performs well when testing is done from a native sputum specimen

Testing a decontaminated specimen has the advantage that the specimen is well homogenized and thus random error is minimal. However, a decontaminated specimen does not pose the same challenges to an assay (in terms of matrix) as a native sputum specimen and does not represent the intended use case defined in the TPP. If comparative testing is done on the pellet, it must be combined with evidence to show that the index test also performs well when testing is done from a native sputum specimen.

We suggest that the approach of first homogenizing a native sputum specimen followed by physical splitting and testing represents a good balance of various considerations, and we recommend this option under most circumstances (see Figure 2). However, beyond the study validity and applicability considerations discussed above, other factors also need to be taken into account when making a choice for the sample flow, eg, feasibility of multiple patient visits, available funding, other available data on index test, etc. Aiming to perform the index test, comparator and reference standard on the same sputum specimen introduces another design challenge, namely that of specimen volume. More specifically, specifying a higher minimum volume requirement as a participation eligibility criterion may allow for more tests to be done on a single specimen, but this may lead to exclusion of patients who cannot produce a high-volume specimen, which in turn can affect generalizability. We recommend that initial studies ensure sufficient volume to allow index test and comparator tests as well as reference tests to be performed on the same specimen. Subsequent studies should include all patients who can provide a specimen (irrespective of volume) to assess accuracy in patients who are only able to provide low-volume specimens.

KEY ISSUES BEYOND ACCURACY

High diagnostic accuracy—and the data demonstrating it—are necessary but insufficient for a test to be supported by policy [31] and to have an impact on patient health, population health, and health system functioning. This explains the additional criteria included in the original TPP [15]. Indeed, although stakeholders rated sensitivity as the most important test attribute in the smear-replacement TPP, stakeholders also focused on the following TPP attributes: simple maintenance/calibration; reagent kit storage/stability; simple specimen preparation steps; and time to results [32]. Other key supporting elements around a test include comprehensive training materials, maintenance and support systems, quality assurance, and connectivity, because policymakers are looking at the practicality of adopting an entire test ecosystem, not simply a single test [33]. Cost, ease of use, and biosafety considerations are also essential components of the TPP and need to be assessed as well, as are other attributes such as infrastructure requirements, availability of other assays to use on the same instrument (for multidisease testing), and an instrument's physical footprint among others. The current article provides guidance for the assessment of test accuracy. Standard approaches are also required to assess other attributes such as biosafety requirements, cost, durability, ease of maintenance, and connectivity—these issues are discussed briefly in the TPP [15], but best practices around their application in study settings need to be further refined.

Another important area (outside of the remit of this piece) is to assess delivery models for new diagnostic tests using implementation research and to assess a new test's potential impact on patient-relevant outcomes through modeling and pragmatic clinical trials [34]. The impact of a new test will vary depending on (1) the existing standard of care for testing, (2) the functioning of the health system in which the test is introduced, and (3) how the new test is implemented [35–37]. For example, empirical therapy partially compensates for the insufficient sensitivity of sputum smear microscopy, at least in patient groups where treatment thresholds are low [38]. In settings where empirical treatment is common, product innovations (such as novel diagnostics more sensitive than smear) may have a less-than-expected impact. Studies evaluating novel smear-replacement tests should report the impact of empirical treatment by including the number of patients diagnosed with TB by the index test (and not the comparator test) who were treated empirically and the time to definitive TB diagnosis. Likewise, delays in TB treatment initiation and/or high rates of pretreatment losses to follow-up may undermine the impact of novel diagnostics. The introduction of new tests may allow changes in how care is delivered (eg, "same-day test-and-treat" may become more feasible than with smear microscopy), but if such changes are not implemented, impact will be blunted. Implementation science research is needed to identify how health systems should best adapt their workflow—linking more sensitive and rapid tests to TB treatment initiation and completion—to realize the full potential of these tests on important clinical and public health outcomes [39, 40].

CONCLUSIONS

This article provides guidance for diagnostic test accuracy studies of sputum smear-replacement tests. We address key study design considerations with regards to the study population, reference standard, use of comparators, and issues related to the complexity of sputum as a specimen matrix. Considering this guidance will (1) facilitate study planning, (2) improve study quality, consistency, and comparability, and (3) ultimately support policy development and scale-up of new smear-replacement tests.

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).

Financial support. S. E. D. is partially funded by the National Institute of Allergy and Infectious Diseases (grant number K24AI104830). G. T. is partially funded by the South African government through the South African Medical Research Council and the EDCTP programme supported by the European Union (project number SF1041).

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

1. Steingart KR, Ng V, Henry M, et al. Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis **2006**; 6:664–74.
2. Berger HW, Mejia E. Tuberculous pleurisy. Chest **1973**; 63:88–92.
3. Scharer L, McClement JH. Isolation of tubercle bacilli from needle biopsy specimens of parietal pleura. Am Rev Respir Dis **1968**; 97:466–8.
4. Huhti E, Brander E, Paloheimo S, Sutinen S. Tuberculosis of the cervical lymph nodes: a clinical, pathological and bacteriological study. Tubercle **1975**; 56:27–36.
5. Thwaites GE, Chau TT, Stepniewska K, et al. Diagnosis of adult tuberculous meningitis by use of clinical and laboratory features. Lancet **2002**; 360:1287–92.
6. Schepers GW. Tuberculous pericarditis. Am J Cardiol **1962**; 9:248–76.
7. Kunkel A, Abel Zur Wiesch P, Nathavitharana RR, Marx FM, Jenkins HE, Cohen T. Smear positivity in paediatric and adult tuberculosis: systematic review and meta-analysis. BMC Infect Dis **2016**; 16:282.
8. Steingart KR, Schiller I, Horne DJ, Pai MP, Boehme CC, Dendukuri N. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database Syst Rev **2014**:CD009593.
9. World Health Organization. Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children: policy update. Geneva: World Health Organization; **2013**.
10. Cazabon D, Suresh A, Oghor C, et al. Implementation of Xpert MTB/RIF in 22 high tuberculosis burden countries: are we making progress? Eur Respir J **2017**; 50:1700918.
11. Boehme CC, Nicol MP, Nabeta P, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. Lancet **2011**; 377:1495–505.
12. Hsiang E, Little KM, Haguma P, et al. Higher cost of implementing Xpert(®) MTB/RIF in Ugandan peripheral settings: implications for cost-effectiveness. Int J Tuberc Lung Dis **2016**; 20:1212–8.
13. Drain PK, Gardiner J, Hannah H, et al. Guidance for studies evaluating the accuracy of biomarker-based non-sputum tests to diagnose tuberculosis. J Infect Dis **2019**; 220(Suppl 3): S108–S16.
14. MacPherson P, Houben RM, Glynn JR, Corbett EL, Kranzer K. Pre-treatment loss to follow-up in tuberculosis patients in low- and lower-middle-income countries and high-burden countries: a systematic review and meta-analysis. Bull World Health Organ **2014**; 92:126–38.
15. World Health Organization. High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28–29 April 2014, Geneva, Switzerland. Geneva: World Health Organization; **2014**.
16. Georghiou SB, Schumacher SG, Rodwell TC, et al. Guidance for Studies Evaluating the Accuracy of Rapid Tuberculosis Drug-Susceptibility Tests. J Infect Dis **2019**; 220(Suppl 3):S127–S36.
17. Foundation for Innovative New Diagnostics. FIND Dx Pipeline Tracker. Available at: https://www.finddx.org/dx-pipeline-status/. Accessed May 2018.
18. Denkinger CM, Nicolau I, Ramsay A, Chedore P, Pai M. Are peripheral microscopy centres ready for next generation molecular tuberculosis diagnostics? Eur Respir J **2013**; 42:544–7.
19. Tuberculosis Coalition for Technical Assistance. International Standards for Tuberculosis Care (ISTC). The Hague: Tuberculosis Coalition for Technical Assistance; **2006**. Available at: https://www.who.int/tb/publications/2006/istc_report.pdf?ua=1. Accessed May 2018.
20. Lewinsohn DM, Leonard MK, LoBue PA, et al. Official American Thoracic Society/Infectious Diseases Society of America/Centers for Disease Control and Prevention Clinical Practice Guidelines: diagnosis of tuberculosis in adults and children. Clin Infect Dis **2017**; 64:111–5.
21. World Health Organization. Guidelines for intensified tuberculosis case-finding and isoniazid preventive therapy for people living with HIV in resource-constrained settings. Geneva, Switzerland: World Health Organization; **2011**. Available at: http://apps.who.int/iris/bitstr

eam/10665/44472/1/9789241500708_eng.pdf. Accesed May 2018.
22. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med **2002**; 137:598–602.
23. Class II special controls guideline: nucleic acid-based in vitro diagnostic devices for the detection of *Mycobacterium* tuberculosis complex in respiratory specimens. Food and Drug Administration. Maryland: Silver Spring; **2014**.
24. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ **2016**; 353:i3139.
25. Denkinger CM, Schumacher SG, Gilpin C, et al. Guidance for the evaluation of tuberculosis diagnostics that meet the world health organization (who) target product profiles: an introduction to who process and study design principles. J Infect Dis **2019**; 220(Suppl 3):S91–S98.
26. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ **2002**; 324:669–71.
27. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med **2013**; 158:544–54.
28. Dorman SE, Chihota VN, Lewis JJ, et al. Performance characteristics of the Cepheid Xpert MTB/RIF test in a tuberculosis prevalence survey. PLoS One **2012**; 7:e43307.
29. Kelly-Cirino CD, Musisi E, Byanyima P, et al. Investigation of OMNIgene·SPUTUM performance in delayed tuberculosis testing by smear, culture, and Xpert MTB/RIF assays in Uganda. J Epidemiol Glob Health **2017**; 7:103–9.
30. Hadad DJ, Morais CG, Vinhas SA, et al. Evaluation of processing methods to equitably aliquot sputa for mycobacterial testing. J Clin Microbiol **2012**; 50:1440–2.
31. Schünemann HJ, Mustafa R, Brozek J, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol **2016**; 76:89–98.
32. Adepoyibi T, Lilis L, Greb H, Boyle D. Which attributes within target product profiles for tuberculosis diagnostics are the most important to focus on? Int J Tuberc Lung Dis **2018**; 22:425–8.
33. Albert H, Nathavitharana RR, Isaacs C, Pai M, Denkinger CM, Boehme CC. Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? Eur Respir J **2016**; 48:516–25.
34. Schünemann HJ, Oxman AD, Brozek JL, et al. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ **2008**; 336:1106–10.
35. Lin HH, Dowdy D, Dye C, Murray M, Cohen T. The impact of new tuberculosis diagnostics on transmission: why context matters. Bull World Health Organ **2012**; 90:739–47A.
36. Sun AY, Denkinger CM, Dowdy DW. The impact of novel tests for tuberculosis depends on the diagnostic cascade. Eur Respir J **2014**; 44:1366–9.
37. Schumacher SG, Sohn H, Qin ZZ, et al. Impact of molecular diagnostics for tuberculosis on patient-important outcomes: a systematic review of study methodologies. PLoS One **2016**; 11:e0151073.
38. Theron G, Peter J, Dowdy D, Langley I, Squire SB, Dheda K. Do high rates of empirical treatment undermine the potential effect of new diagnostic tests for tuberculosis in high-burden settings? Lancet Infect Dis **2014**; 14:527–32.
39. Pai M, Schumacher SG, Abimbola S. Surrogate endpoints in global health research: still searching for killer apps and silver bullets? BMJ Glob Health **2018**; 3:e000755.
40. Schumacher SG, Thangakunam B, Denkinger CM, et al. Impact of point-of-care implementation of Xpert® MTB/RIF: product vs. process innovation. Int J Tuberc Lung Dis **2015**; 19:1084–90.