Stop TB Partnership
New Diagnostics Working Group

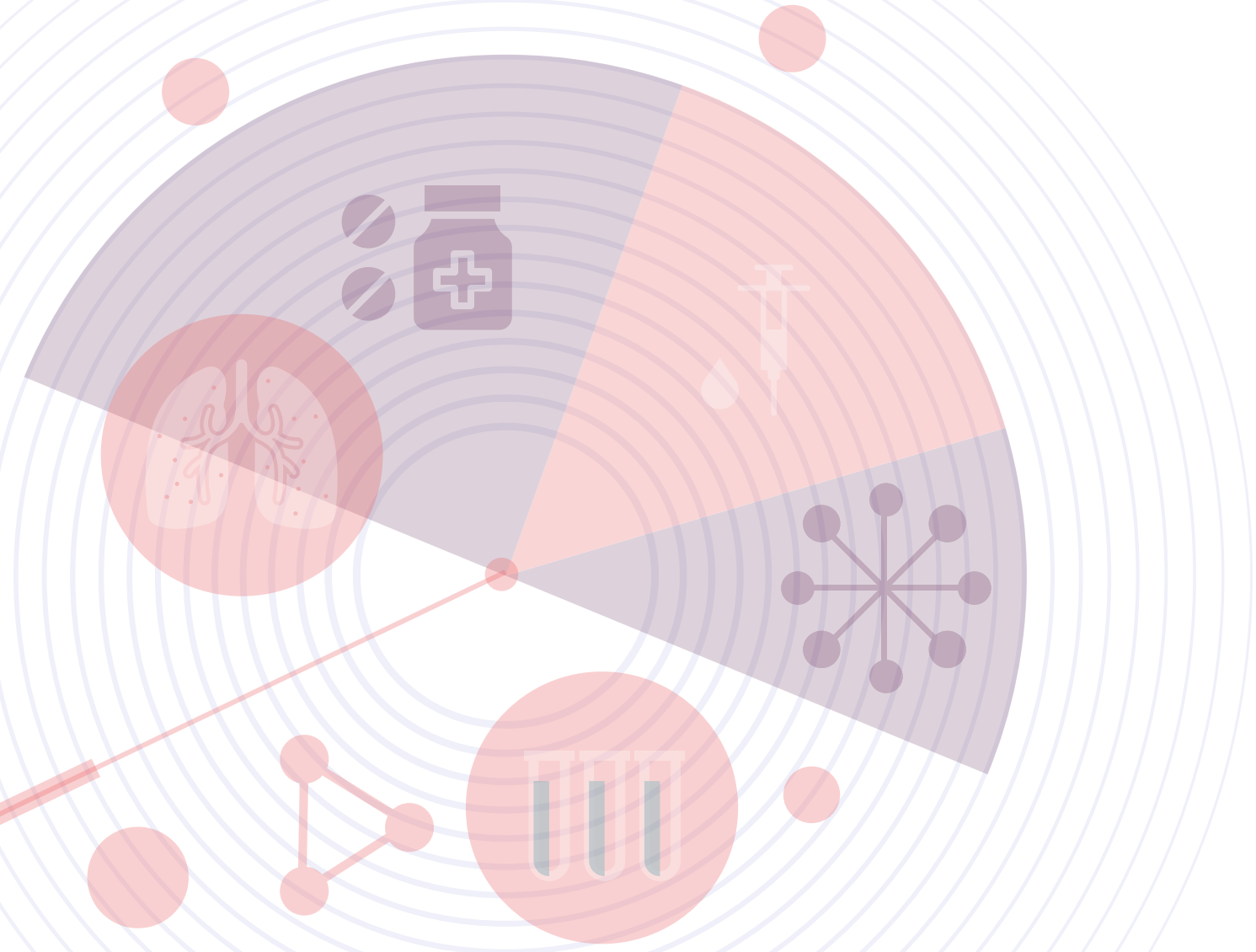
World Health Organization

Framework for the evaluation of new tests for **tuberculosis infection**

Stop TB Partnership
New Diagnostics Working Group

World Health Organization

Framework for the evaluation of new tests for **tuberculosis infection**



Framework for the evaluation of new tests for tuberculosis infection

ISBN 978–92–4-000718–5 (electronic version)
ISBN 978–92–4-000719–2 (print version)

Suggested citation. Framework for the evaluation of new tests for tuberculosis infection. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see http://apps.who.int/bookorders. To submit requests for commercial use and queries on rights and licensing, see http://www.who.int/about/licensing.

Editing and design by Inís Communication

Contents

Acknowledgements

This document was prepared by Yohhei Hamada (Research Institute of Tuberculosis, Japan) and Alberto Matteelli (University of Brescia, Italy), with input from a technical expert group comprising the following participants:

New Diagnostics Working Group, Stop TB Partnership, Geneva, Switzerland: Daniela Cirillo, Alberto Matteelli, with support from Karishma Saran in her capacity as Secretariat, New Diagnostics Working Group.

Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland: Adam Penn-Nicholson, Morten Ruhwald.

McGill University, Montreal, Canada: Dick Menzies, Olivia Oxlade.

World Health Organization Global TB Programme, Geneva, Switzerland: Saskia den Boon, Dennis Falzon, Avinash Kanchar, Alexei Korobitsyn, Matteo Zignol.

The technical working group used a consensus-building process to develop the document.

The participants in the external review group were Sevim Ahmedov, Claudia Denkinger, Darragh Duffy, Nazir Ismail, Afrânio Kritski, Elisa Nemes and Molebogeng X. Rangaka.

Abbreviations

ART	antiretroviral treatment
BCG	bacille Calmette–Guérin
CV	coefficient of variation
CXR	chest radiography
ELISA	enzyme-linked immunosorbent assay
HIV	human immunodeficiency virus
IGRA	interferon-gamma release assay
LLoQ	lower limit of quantification
LoD	limit of detection
PLHIV	people living with HIV
SD	standard deviation
TB	tuberculosis
TPT	TB preventive treatment
TST	tuberculin skin test
WHO	World Health Organization

Background

About a quarter of the world's population is estimated to be infected with *Mycobacterium tuberculosis* *(1, 2)*. People with the infection are at risk of progressing to active tuberculosis (TB) disease. The lifetime risk of developing TB among people infected with *M. tuberculosis* is 5–15%, with a peak in the first 2 years after infection (recent infection). The risk of progression varies between individuals and is influenced mainly by host factors such as comorbidities, age and nutritional status *(3)*. The risk is higher in people who have been recently infected (e.g. contacts of people with TB).

Treatment of TB infection,[1] also known as tuberculosis preventive treatment (TPT), is one of the critical components to achieving the ambitious targets of the End TB Strategy 2016–2035 *(4)*. At the first United Nations High-level Meeting on TB in 2018, Member States committed to provide TPT to at least 30 million people in 2018–2022, including 6 million people living with human immunodeficiency virus (HIV), 4 million children aged under 5 years who are household contacts of people with TB, and 20 million other household contacts *(5)*. There is no gold-standard method for diagnosing TB infection *(3)*. The World Health Organization (WHO) currently recommends a tuberculin skin test (TST) or an interferon-gamma release assay (IGRA) to test for TB infection to identify suitable candidates for TPT *(6)*. These tests measure immune sensitization by *M. tuberculosis*; however, a positive test does not necessarily indicate the presence of living TB bacilli, and so tests can remain positive despite an adequate course of TPT. These tests are helpful to identify people at higher risk of developing TB, because in most published studies such risk is higher in people who test positive for TB infection than in those who test negative *(7)*. Current tests for TB infection have limited value, however, in predicting the risk of progression from infection to TB disease *(8)*: less than 10% of people with a positive test for TB infection develop disease over a 2-year period *(7, 9)*.

The development and evaluation of tests characterized by higher prediction capacity (tests of progression) is a high priority for research. In 2017 WHO published target product profiles for such tests *(10)*. These tests should be able to identify people who are likely to develop TB disease within the subsequent two years and, unlike TST or IGRA, should provide negative results in people who do not progress to active disease. According to the target product profiles defined by WHO, the optimal sensitivity and specificity of such tests for predicting development of TB disease are at least 90% *(10, 11)* – much higher than those of TST or IGRA *(12)*. Until such tests become available for use under field conditions, tests for TB infection, including TST and IGRA, remain the standard tests of choice.

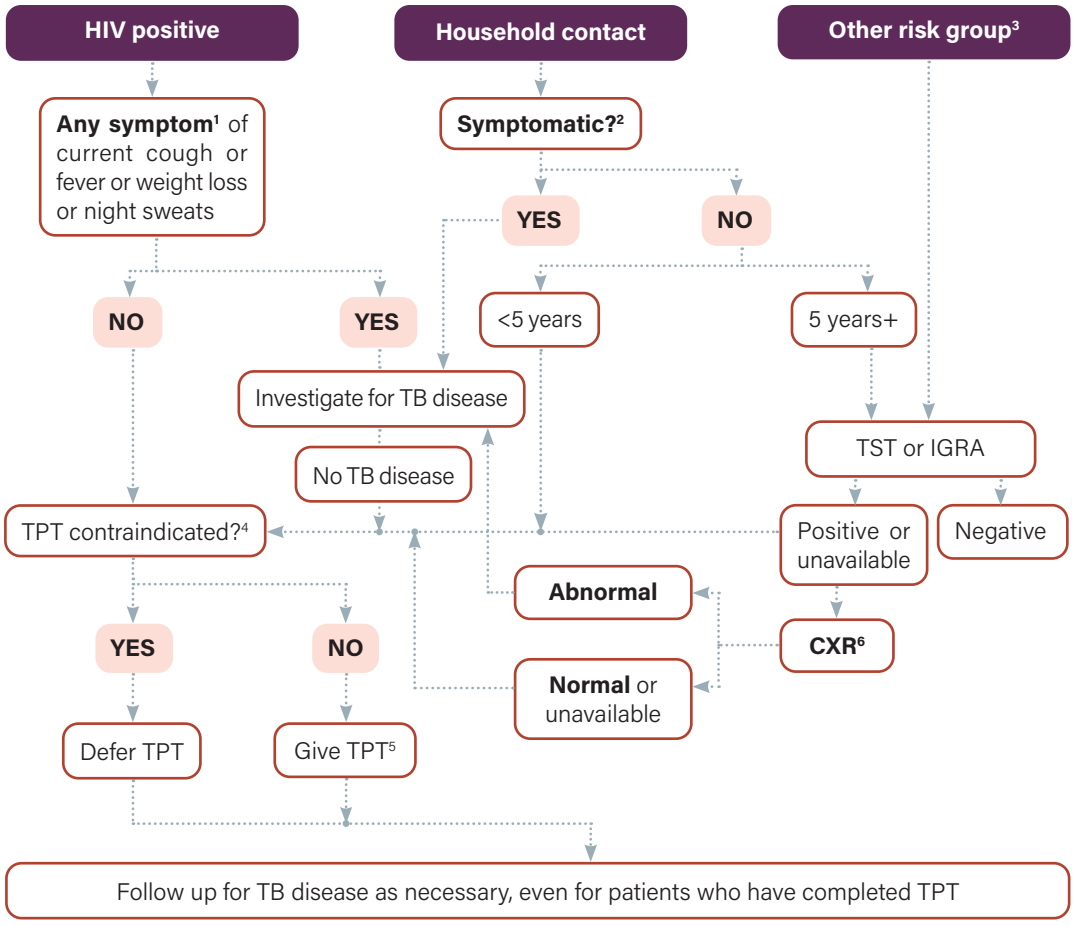
Partly as a result of testing availability and limitations in their accuracy, tests for TB infection are not required before starting TPT in people from high-priority groups considered to be at risk in high-burden countries, such as people living with HIV and household contacts aged under five years (Figure 1) *(6)*. For people from other at-risk populations, tests for TB infection are recommended to identify those who would benefit most from treatment and to avoid unnecessary medication (Figure 1). However, implementation of tests for TB infection is fraught with difficulties, including high

[1] Given that infection cannot always be considered latent and that the main difference between active and latent TB is the presence or absence of disease, unless otherwise stated we use the term TB infection to represent all stages of infection with *M. tuberculosis* without clinical manifestations of TB disease.

costs (IGRA), cold-chain requirements (TST), short supply of quality-assured TST, and inadequate laboratory set-up to undertake high volumes of IGRA testing in decentralized settings. This calls for new tests with better operational characteristics.

In 2018, the year after the WHO guidelines first recommended TPT in all household contacts regardless of setting, fewer than 80 000 household contacts aged 5 years or over were reported to have been started on TPT globally *(13)* – far below the average of 4 million per year needed to achieve the minimum target set by countries in the Political Declaration at the United Nations High-level Meeting on TB in 2018 *(5)*. Household contacts aged 5 years and over represent two-thirds of the 30 million target set by the High-level Meeting for 2018–2022. This adds urgency to the need to accelerate the scale-up of testing for TB infection and to find better-performing tests in the near future.

Figure 1. Algorithm for testing and treating tuberculosis infection in different groups considered to be at risk



1. If <10 years, any one of current cough or fever or history of contact with TB or reported weight loss or confirmed weight loss >5% since last visit or growth curve flattening or weight for age <-2 Z-scores. Asymptomatic infants <1 year with HIV are only treated for LTBI if they are household contacts of TB. TST or IGRA may identify PLHIV who will benefit most from preventive treatment. Chest radiography (CXR) may be used in PLHIV on ART, before starting LTBI treatment.
2. Any one of cough or fever or night sweats or haemoptysis or weight loss or chest pain or shortness of breath or fatigue. In children <5 years, they should also be free of anorexia, failure to thrive, not eating well, decreased activity or playfulness to be considered asymptomatic.
3. Including silicosis, dialysis, anti-TNF agent treatment, preparation for transplantation or other risks in national guidelines.
4. Including acute or chronic hepatitis; peripheral neuropathy (if isoniazid is used); regular and heavy alcohol consumption. Pregnancy or a previous history of TB are not contraindications.
5. Regimen chosen based on considerations of age, strain (drug susceptible or otherwise), risk of toxicity, availability and preferences.
6. CXR may have been carried out earlier on as part of intensified case finding.

Source: WHO consolidated guidelines on tuberculosis: tuberculosis preventive treatment. Geneva: World Health Organization; 2020.

New versions of TST and IGRA are expected to become available in the near future, all using recombinant ESAT6 and CFP10 antigens. New skin-based tests for TB infection include C-Tb (Serum Institute of India, India), Diaskintest (Generium, Russian Federation) and ESAT6-CFP10 (Anhui Zhifei Longcom, China). Qiagen (the Netherlands), the manufacturer of the IGRA test QuantiFERON-TB Gold Plus, and SD Biosensor (Republic of Korea) have both developed simplified versions of IGRA that can be operated in peripheral facilities without laboratory infrastructure. These new versions might allow easier identification of people eligible for TPT.

The development and evaluation of novel tests for the identification of people who should receive TPT is a priority in achieving the targets of the End TB strategy and in countries aiming to eliminate TB. Evaluation of tests for TB infection is not straightforward due to the lack of a gold standard. Guidance on evaluation of tests for TB infection may facilitate a standardized evaluation of new tests for TB infection and accelerate adoption into global and national policies and subsequent scale-up.

Scope

The main aim of this document is to provide test manufacturers, researchers, research funders, regulators, TB programme coordinators, civil society and other stakeholders with a framework for evaluation of new immunodiagnostic tests for TB infection. In this document, tests for TB infection such as TST and IGRA are differentiated from tests for progression or tests for incipient TB, which are intended to predict progression from TB infection to TB disease. Once endorsed, such tests would have advantages in the selection of people who would benefit from TPT. A framework to evaluate tests for progression is described elsewhere *(10, 11)*.

The scope of the document is to describe the principles to be considered when evaluating new tests for TB infection. It aims to promote and direct research by identifying standard study designs and evaluation protocols. The document also provides guidance on the operational and performance characteristics of tests for TB infection. It does not aim to be prescriptive or to provide details on thresholds for analysis or minimum requirements for diagnostic performance of new tests for TB infection.

Document development process

In November 2019 WHO established a technical expert group to provide guidance on the framework for evaluation of tests for TB infection. The technical expert group was a partnership between the WHO Global TB Programme and the New Diagnostics Working Group of the Stop TB Partnership, the Foundation for Innovative New Diagnostics (FIND) and McGill University. Members of the technical expert group were selected based on their subject expertise. The group developed a draft document through a series of in-person and virtual meetings. A non-systematic review of previous studies and protocols for evaluating tests for TB infection informed the development of the document. The draft was circulated through the members of an ad hoc expert review committee before being finalized.

Framework for evaluation of new tests for tuberculosis infection

Considerations for evaluation of new tests

The most important aspects to be considered in studies evaluating new tests for TB infection are:

- diagnostic and predictive performance;
- laboratory characteristics, including analytical sensitivity (limit of detection, limit of quantification), precision, repeatability and reproducibility, and stability;
- safety for skin tests;
- costs to the health system and individuals;
- operational characteristics.

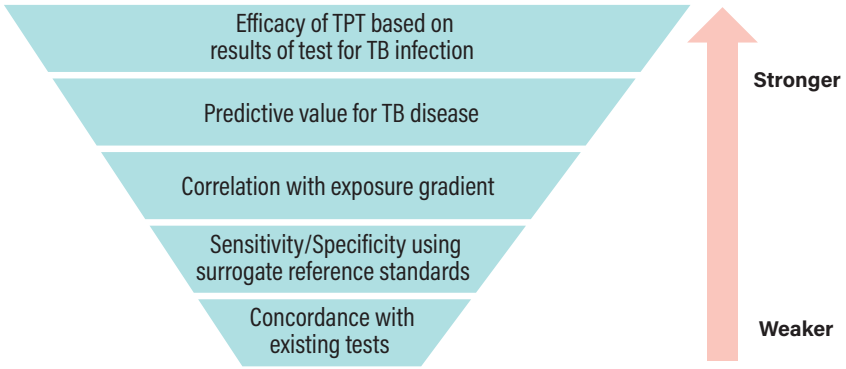
Guidance on studies evaluating diagnostic performance of tests for TB infection, including study design, population, procedures, analysis and sample size calculation, is given below. Technical issues and evaluation of safety for skin tests are also discussed.

New tests for TB infection, including C-Tb, Diaskintest and ESAT6-CFP10, and point-of-care IGRA tests, including QuantiFERON Access, Standard E and F TB-feron (SD Biosensor, Republic of Korea) and Advansure TB-IGRA (LG Chem, Republic of Korea), are not expected to provide major advantages in terms of accuracy or predictive ability. These tests can, however, offer significant advantages in terms of operational characteristics, such as feasibility in low-resource settings or reduced costs to the health system and patients. These aspects are also important considerations when evaluating tests for TB infection and thus also addressed in this document.

Hierarchy of reference standards

Studies evaluating the performance of TB infection are hampered by the lack of an adequate gold standard to distinguish the presence or absence of TB infection. As a result, the estimation of sensitivity and specify, which are usually standard measures for assessing performance of diagnostic tests, require surrogate reference standards (see below). Hence, a hierarchy of reference standards was developed when WHO reviewed evidence on the use of IGRA (Figure 2) *(12)*.

Figure 2. Hierarchy of reference standards used to assess the evidence base for tests for tuberculosis infection



Adapted from: Use of interferon gamma release assays (IGRAs) in tuberculosis control in low-and middle-income settings. Geneva: World Health Organization; 2010 Expert Group Meeting Report. 20-21 July 2010. WHO/HTM/TB/2011.17

Reference standards at higher levels of hierarchy give a stronger evidence base for a test's predictive performance to identify people who would benefit from TPT. Any new test for TB infection will likely be based on similar concepts to TST and IGRA – that is, determining an immune response to *M. tuberculosis*-specific antigens in vivo (size of skin induration) or in vitro (magnitude of cytokine release) – and hence they are not expected to offer significant improvement in predictive performance. In general, any new test for TB infection should have predictive performance at least as good as (not inferior to) the currently available tests (although their predictive performance is low). On the other hand, it is challenging to conduct a study to measure predictive value, as detailed below. Therefore, sensitivity, specificity and concordance may be a primary endpoint for comparative studies of new tests for TB infection. In this case, a new test for TB infection could demonstrate non-inferior sensitivity and specificity or concordance compared with at least one of the currently available tests endorsed by WHO as rule-in tests for TPT (TST or IGRA).[2] Because there is mounting evidence that IGRA has higher specificity and possibly higher sensitivity than TST, IGRA should be preferred as a comparator in new trials *(14, 15)*.

Any new tests intended to achieve significant improvement in predictive performance should follow the evaluation framework for tests for predicting progression to TB disease *(10, 11)*.

Study design and population

Study design 1: predictive performance

A prospective longitudinal study measuring predictive value is the most appropriate method to compare performance of tests for TB infection. In the WHO TPT guidelines, it is recommended that either TST or IGRA is used to test for TB infection (strong recommendation, very low-quality evidence) *(10)*. The equivalence of TST and IGRA was based on the results of a systematic review of 29 studies on the predictive utility of TST and IGRA *(16)*. In this study design, people with positive or negative tests for TB infection are screened for TB disease; those who are free of TB disease are then followed for at least 12 months for development of TB disease. The major ethical issue with this design is that eligible people with a positive test for TB infection should be recommended, and encouraged, to take TPT, which reduces TB incidence; hence it is not possible to estimate predictive performance of the

[2] Two commercially available IGRA (QuantiFERON®-TB Gold In-Tube, T-SPOT®.TB) are currently endorsed by WHO. Later versions expected to have better performance (e.g. QuantiFERON-TB Gold-Plus) will be evaluated by WHO.

test. Although incidence of disease can be measured in people who test positive for TB infection but who do not take treatment, these people often have very different characteristics from those who do take treatment. This may bias the estimates of the predictive ability. In addition, people who do not follow provider recommendations may also be more difficult to follow prospectively. If the number of people lost to follow-up is high, this may seriously jeopardize the integrity of the cohort design. However, if two or more tests for TB infection are tested in the same cohort, then the relative incidence rates of disease between people who tested positive to the different tests for TB infection but who do not take treatment provides a relative (not absolute) estimate of predictive ability. Detailed guidance is available in the framework for evaluation of tests for progression *(10, 11)*.

Study design 2: sensitivity and specificity using clinical reference standards

Sensitivity can be assessed using the clinical standard of TB disease, which is confirmed microbiologically by culture. Using microbiologically confirmed TB as a reference standard is particularly important and should be encouraged for paediatric TB, which tends to be diagnosed clinically and whose accuracy is sometimes questionable. When that is not feasible, the diagnosis can exceptionally be based on other clinical criteria, such as radiographic criteria, particularly for extrapulmonary TB or for TB in children aged under five years. In this case, it is important not to include tests for TB infection as part of the diagnosis, as this leads to incorporation bias, which will substantially overestimate the sensitivity.

Specificity depends on the antigens used to stimulate the memory T-cells. By using overlapping peptides from highly *M. tuberculosis*-specific antigens such as ESAT-6 and CFP-10, current IGRA tests are, unlike TST, not affected by bacille Calmette–Guérin (BCG) vaccination and their specificity is greater than 95% in low-risk populations. However, expression of ESAT-6 and CFP-10 by a group of nontuberculous mycobacteria species, including *M. marinum* and *M. kansasii*, is well known to lead to false-positive test results in people infected with these species. For estimation of specificity, the ideal population is the one with very low likelihood of prior exposure to *M. tuberculosis*. It is important to evaluate the impact of cross-reactions by conducting subgroup analyses by BCG vaccination status and by likelihood of exposure to nontuberculous mycobacteria. For assessment of tests for TB infection based on TB-specific antigens not found in the BCG vaccine itself (such as ESAT-6 or CFP-10), BCG vaccination is not relevant in population selection.

Participants should be selected randomly or consecutively when enrolled. It is particularly important to avoid selecting the study population from a larger group of potentially eligible people on the basis of clinical or disease characteristics that might affect test performance (although this is acceptable in early stages of the evaluation of new tests). The key concept is to include participants who represent, as closely as possible, the populations in which the tests will be used.

The fundamental study design is cross-sectional, meaning the tests are performed at the same time as the clinical evaluation is made. This can be part of a cohort study, but longitudinal follow-up is not essential. It is also important to conduct the tests in different TB epidemiological settings, although specificity cannot be measured in settings with high prevalence of TB infection.

The tests should be performed by well-trained people experienced in the procedures for both the reference test and the new test. People performing and evaluating tests should be blinded to the results of the other tests. Methods for supervision, monitoring and quality control should be adequate and clearly described. These points also apply to studies of agreement.

Study design 3: concordance of tests

This design is essentially a study of agreement between new and reference tests and is considered to be the lowest level of evidence for assessment of diagnostic tests. It should be noted that concordance is not expected when new tests are thought to have superior predictive performance or sensitivity and specificity. This design is therefore appropriate when the new test offers operational advantages over existing tests but no gain in diagnostic performance is expected.

Study participants should be representative of the general population. For example, it is important to avoid excluding very young or elderly people, people with more severe disease, or people with serious comorbidities (e.g. HIV, diabetes, renal failure, malnutrition), as these may affect test performance. Additionally, by including contacts with different grades of exposure, it is possible to evaluate the correlation between positivity and the level of exposure, which will give more confidence in the value of the test. It is encouraged to study agreement in different TB epidemiological settings and expected burdens of nontuberculous mycobacteria.

When two tests are administered, they should be done at the same time, ideally on the same sample (if based on blood or urine). If analysing skin tests, the test should be administered on two different sites at the same time. If the tests cannot be given at the same time, then the interval between the two tests should be minimized and be within an interval considered unlikely to differentially affect the performance of tests. Collection of blood or urine specimens for in vitro tests for TB infection should be done at the same time as or before TST to avoid a boosting effect.

Sample size calculation

Principles of sample size determination: superiority versus non-inferiority

Superiority designs and related sample size calculations are appropriate if the reference test has suboptimal performance. For example, the sensitivity of current tests for TB infection is judged to be suboptimal when tested in the population with TB disease – a surrogate for TB infection, but the only situation when we can be certain that TB infection has occurred. Sensitivity is particularly suboptimal in young children, people living with HIV, and otherwise immunocompromised people, who are at increased risk of disease and therefore a priority for testing for TB infection. Therefore, a new test would be of great interest if it had superior sensitivity, particularly in these high-risk populations.

If the reference test performance is excellent, then it may be almost impossible to demonstrate superiority. For example, the current IGRA tests have excellent specificity, in the range of 97–99%. Given this, it would require a huge sample size to demonstrate superiority. However, if the specificity of a new test is "not significantly worse", then it is desirable for the new test to have other advantages. If a non-inferiority design is selected, then these other important advantages, such as lower cost, enhanced feasibility or point-of-care availability, must be prespecified and measured carefully. It would also be recommended to perform power calculations for these other outcomes.

Non-inferiority designs prespecify that a new test or intervention will be considered acceptable if it is "not significantly worse" – but because of the possibility of a type II error, the new test may be worse than the reference test and yet still declared non-inferior. If a subsequent study is done demonstrating non-inferiority of a third new test to this newly adopted standard reference, then we could see a gradual creep towards inferior performance being considered acceptable. Hence, non-inferiority designs should be reserved for aspects of test performance where the standard or reference test truly has very good performance, such as the example of the specificity of IGRA tests.

General determinants of sample size

When designing a study, it is very important to prespecify the effect size, or difference expected, for both superiority and non-inferiority designs. In general, the larger the effect size, the smaller the resultant sample size (see Annex 1 for examples of how changes in effect sizes can result in changes in the required sample size). It is important to recognize that larger effect sizes may not be realistic and may lead to erroneous conclusions. For superiority designs, if the new test is hypothesized to be superior to the reference test by a wide effect size, then a relatively small number of participants will be needed. In a study with too few participants, a new test that has superior performance but by a smaller effect size may fail to show superiority due to a wide confidence interval crossing the null value, which will be erroneous.

The same is true for non-inferiority. In order to claim non-inferiority, the lower bound (or sometimes upper bound) of the confidence interval of the effect size should not cross the non-inferiority margin. The sample size required for a large non-inferiority margin is much smaller. However, this could lead to the new intervention or test being declared "non-inferior" even if it is minimally inferior *(17, 18)*.

In general, we need to know the reference test performance, as this affects the sample size (see Annex 1). This should be based on recently conducted high-quality systematic reviews to ensure the estimates of the reference test performance are accurate. If local estimates of test performance are different, then it is advisable to perform two sample size calculations – one based on estimates from systematic reviews, and another based on local test performance. It is strongly advised to use the larger of the two sample size calculations in order to answer the question adequately.

Specific sample size calculations

Predictive performance

Reference tests (IGRA or TST) have relatively low sensitivity and very low specificity for predicting TB disease. In most cohort studies, the incidence of TB disease in high-risk individuals tested and not treated is 1–2% in the first 2 years. The new test would ideally have good sensitivity (predict all people with future TB disease) and reasonable specificity (not identify too many people with positive tests who do not develop TB in the future). However, new tests for TB infection are not expected to improve predictive value substantially, and most gains are expected to be seen in operational aspects. Hence, demonstrating non-inferiority in terms of predictive performance would be acceptable.

Sample size calculations have to account for the likelihood of future TB disease, as this determines the number of events, the sensitivity of current tests in predicting these events, and differences in sensitivity of the new test and follow-up time. If the annual event rate is 1–2% per year, then the sample size must be inflated by 50–100 times, and then inflated to account for losses during follow-up. If follow-up is for two years, this reduces the required sample size by about half. Longer periods of follow-up may result in greater losses to follow-up. A large number of people lost during follow-up and with unknown outcomes can jeopardize the integrity of the study findings. Additionally, longer periods of follow-up lead to higher risk of reinfection in high-transmission settings, which could complicate interpretation of the initial test. Hence, longer periods of follow-up are not encouraged – at least for the primary analysis and sample size calculations. The sample size calculations do not make any assumptions on the repeatability/reproducibility of the tests beyond what is subsumed in the sensitivity or specificity considerations.

Sensitivity

Current tests for TB infection generally have suboptimal sensitivity, ranging in published systematic reviews from 72% to 80% in people with TB disease. Sensitivity is lower in children, people living with HIV, and other immunocompromised people. In these populations, superiority of a new test would be preferable to a non-inferiority design. For examples of the effects of different assumptions regarding reference test sensitivity and the differences we want to detect, see Annex 1.

Specificity

Current IGRA tests have excellent specificity, in the range 97–99% (TST specificity in populations that have not been BCG-vaccinated is similar). Therefore, superior specificity is neither necessary nor likely to be demonstrable, except with a huge sample size. Hence, a non-inferiority design is sufficient for specificity when the reference standard test is an IGRA, or a TST in a population that has not been BCG-vaccinated. If the reference standard is TST in a BCG-vaccinated population, then a superiority design is recommended.

Concordance

The sample size is determined by the maximum acceptable width of the kappa 95% confidence interval, the underlying true proportion of positives, and the anticipated value of kappa *(19)*.

Study analysis

Predictive performance

The event rate in cohort studies is typically calculated as the number of events per 100 person-years of follow-up, which accounts for variable follow-up times in a large-scale cohort. Since the same people have two or more tests for TB infection, then the differences in event rates can be directly calculated either as a risk difference:

(1) new test event rate – reference test event rate

or as a risk ratio.

If all people are followed, then incidence rate ratios can be calculated and 95% confidence intervals calculated as proportions. The incidence rate ratio can be estimated as:

(2) incidence rate among people who test positive/incidence rate among people who test negative

If all individuals with positive and negative tests are followed prospectively, then the sensitivity of the tests for the development of TB disease can be estimated as:

(3) number of people who develop TB disease who tested positive/total number of people tested who develop TB disease during the follow-up time

Sensitivity and specificity

Sensitivity is calculated as:

(4) number of people who test true-positive/(number of people who test true positive + number of people who test false negative)

This requires the identification of people with TB originating from the cohort during the follow-up period, as follows:

(5) number of people who test positive/number of people tested who had TB disease

Specificity is calculated as:

(6) number of people who test true negative/(number of people who test false positive + number of people who test true negative)

In a population with very low TB prevalence, specificity can be approximately calculated as:

(7) all people who test negative/all people tested in the very-low-prevalence population

95% confidence intervals can be calculated for a proportion.

Concordance of tests

If the new test is anticipated to have similar diagnostic accuracy to the reference test but has operational advantages, such as low cost or greater feasibility, then tests showing high agreement are valuable. The kappa statistics should be calculated, with the accompanying 95% confidence interval. Kappa statistics account for chance-corrected agreement, which is very important when prevalence of positive tests is either very low or very high. A kappa statistic greater than 0.60 indicates very good agreement, and a kappa statistic greater than 0.80 indicates near-perfect agreement.

Technical issues

Immunoassays are complex assays influenced by multiple sources of variability that can impact results. Table 1 lists a range of typical sources of variability in IGRA-like tests that should be prioritized by developers in the technical description of the assay.

Table 1. Typical sources of variability in interferon-gamma release assay-like tests

Sources of variability	Item	Suggested documentation	Importance
Factors impacting stimulation assay	Blood collection tubes: within- and between-lot variability	Blood drawn from representative people with TB infection, with responses spanning dynamic range of assay and including critical areas around cut-off point A representative number of blood samples from the same person should be drawn and assessed in parallel – e.g. ≥ 5 tubes from the same lot and from 5 individual lots The following are commonly considered acceptable: ▪ Within-lot variability CV ≤ 10% ▪ Between-lot variability CV ≤ 15%	High
	Delay in blood processing and incubation time	Impact of delay from blood-draw to analysis should be described in a representative sample of test-positive individuals, e.g. at 0, 2, 6, 12 hours Incubation time should be described	Medium
	Volume of blood	If test depends on equipment influenced by e.g. air-pressure changes with altitude such as vacutainer tubes, impact of +/– 20% volume change should be assessed in a representative sample of test-positive individuals	Medium
Analytical range of readout assay	LoD	Approximately 20 repetitions of zero standard over multiple assays (e.g. ELISA plates) or multiple tests on other diagnostic platforms (LOD is often expressed as the mean + 3 SD)	High
	LLoQ	Serial dilutions of low standard to approximate LLoQ analysed over multiple assays ($n \geq 3$) to generate precision profile	High
		LLoQ is commonly expressed as lowest concentration from profile that can be measured with < 20% imprecision and inaccuracy	
Imprecision of the readout assay	Intra-assay and inter-assay imprecision	Assessed with representative quality control samples from range of responses seen in samples from people with TB infection It is important to select some samples in assay extremes to cover responses in cut-off point range Samples should be assessed in ≥ 5 independent determinations for each, over each of 5 days The following are commonly considered acceptable: ▪ Intra-assay CV ≤ 10% ▪ Inter-assay CV ≤ 15% (20% at LLoQ)	High
Accuracy of readout assay	Recovery	Spike of recombinant/purified analyte is added to ≥ 3 independent pools of appropriate matrix (e.g. plasma) at 3 different concentrations; acceptable recovery is 80–120%	Medium
		Evaluation of suitable reference materials if available (≥ 5 determinations over 3 concentrations; < 20% imprecision and inaccuracy)	

Sources of variability	Item	Suggested documentation	Importance
Analytical specificity of readout assay	Cross-reactivity	Identified proteins with homology to analyte are spiked (recombinant/purified forms) into independent samples ($n \geq 2$) at 2 concentrations spanning pathophysiological cross-reactant range (if known)	Low
	Parallelism/ dilution linearity (normal working dilution and ≥ 3 serial dilutions of ≥ 3 samples)	Assessed by back-calculating diluted concentration of 4 dilutions to actual concentration, with acceptability limit of ≤ 15%	Low
	Common interferents (e.g. rheumatoid factor, lipids, bilirubin, complement, haemolysate)	Recombinant analyte is spiked into surplus diagnostic samples ($n \geq 3$) with known moderate and high interferent concentrations and recovery is calculated Alternatively, stock interferents can be purchased and spiked into samples with known amounts of analytes Final concentrations are 50 µg/ml or 150 µg/ml bilirubin (conjugated and unconjugated, respectively) or 30 mg/ml triglycerides For testing effects of haemolysis, samples containing known concentrations of analyte can be spiked with haemolysate to produce 5 mg/ml haemoglobin for serum and plasma samples	Medium
	Evaluation of curve-fitting model (≥ 5 determinations over multiple runs)	Imprecision (< 10%; 20% at LLoQ)	Medium
		Inaccuracy (< 10%; 20% at LLoQ) (> 80% of non-zero standards, including highest and lowest, must pass)	
Additional assessments	Inter-laboratory imprecision (reproducibility)	Reference panel of samples analysed in several laboratories	Medium
	Analyte stability	Freeze–thaw stability assessed by determining concentration of biomarker in panel of representative samples freeze–thawed 1–10 times	High (depending on claim in instructions for use)
		Short-term bench stability assessed by determining concentration of biomarker in panel of representative samples left on benchtop for ≤ 48 hours	
		Long-term storage stability (length and temperature)	

CV, coefficient of variation; ELISA, enzyme-linked immunosorbent assay; LLoQ, lower limit of quantification; LoD, limit of detection; SD, standard deviation.

As immunoassays detect responses on a continuous scale, which is converted to a binary outcome as positive or negative by use of a threshold value or cut-off, a description of the variability around this cut-off is of particular relevance. To determine the degree of variability around the cut-off threshold of any new test, we recommend studies are planned using an adequate number of non-symptomatic participants with positive and negative IGRA values representative of normal physiological ranges in cohorts where prior infection is likely and reinfection events are rare. Such populations may include recent adult migrants or health-care workers in low- to middle-endemic countries where TPT is not routinely initiated for a positive IGRA result. Contacts of people with TB who have recently converted to a positive test could also be followed up to assess the possibility of reversion. We recommend longitudinal sampling of at least two serially collected samples collected four weeks apart. Samples should be used to estimate the rate of IGRA conversions/reversions using the predefined cut-off for assay positivity. In cases of IGRA conversion, a third sample may be collected to evaluate whether conversion using the predefined cut-off was sustained due to an *M. tuberculosis* infection event rather than a spurious effect of variability around the cut-off. Samples from such participants should be excluded when defining the range of the zone of uncertainty.

Given the uncertainty around the cut-off thresholds, assays that report only IGRA results as a binary outcome (positive/negative) may fail to provide confidence in the assessment by a clinical worker. Therefore, we encourage manufacturers to consider providing data to the operator on the magnitude of the response and the classification status.

The QuantiFERON-TB Gold and QuantiFERON-TB Gold Plus assays report considerable variability around the assay cut-off of 0.35 IU/ml. Although the cut-off value for any new IGRA should be defined against the concordance with existing tests such as QuantiFERON-TB Gold, manufacturers should be cautious about reporting test results around this zone of uncertainty. For example, new test data can be reported in comparison to QuantiFERON-TB Gold at defined cut-off thresholds:

- high-certainty IGRA-negative: ≤ 0.2 IU/ml
- low-certainty IGRA: > 0.2 to ≤ 0.7 IU/ml
- high-certainty IGRA-positive: > 0.7 IU/ml
- all IGRA results: independent of IU/ml score.

Assessment of any new IGRA should also report IU/ml for the mitogen, the antigen and the unstimulated control separately. Each new IGRA will need to clearly define the zone of uncertainty within their own assay. Assays with only a binary readout for TB infection should provide additional data to confirm reproducibility without resulting in a high number of invalid results. Participant samples from both uninfected controls from low-endemic settings and controls with *M. tuberculosis* infection should be used to establish a range of interferon-gamma responses expected for clinically relevant specimens.

Ideally, initial studies on immunological tests performance should be conducted in accredited settings where standard operating procedures for good laboratory practice are in place and human blood is safely transported, handled and processed. There is no need to use biosafety level 2 or 3 facilities. The required equipment will depend on the degree of automation of the technology proposed.

If the study is conducted using QuantiFERON-TB Gold as a comparator, basic equipment (centrifuge with closed buckets, incubator, ELISA washer and reader) will be needed. It is mandatory that the people performing the comparator test are fully trained and proficient to minimize preanalytical and analytical variability, while also being representative of the staff in settings where the tests are

intended to be used. Training should be performed by the manufacturer. Laboratories using any IGRA may consider implementing an internal quality assurance programme to monitor intralaboratory performance of IGRA tests over time and across different manufacturing lots. This will be particularly important where research evaluating serially collected IGRA samples is involved.

External monitoring should be performed during the study, recording any variation from the protocol.

Evaluation of safety for skin tests

Safety of new skin tests should be evaluated against a reference skin test (TST) in a population representative for the target population for the new test to show that the number of injection site reactions and other adverse events is similar to or fewer than that seen with TST. Safety should be evaluated in various groups, such as people living with HIV, children, and pregnant and lactating women. The study design should be adequate and seek to minimize bias from interference from prior skin testing, BCG vaccination, or concomitant medication known to impact immune responses. One method to avoid interference from prior skin testing is to give the new test and the reference test at the same time in each forearm in a double-blinded manner. Another method to assess safety is to compare adverse events reported in a randomized controlled trial; however, trials are unlikely to be sufficient to detect rare adverse events, and post-marketing surveillance is essential.

Local and systemic adverse events should be assessed at relevant time points through medical assessments, such as 30 minutes after the skin test injections, on the day of test reading and after 1 month. Methods for recording adverse events should be adequate and clearly recorded, for example using Medical Dictionary for Regulatory Activities classification.

As skin tests are designed to induce a local reaction, it is important to predefine how to interpret reactions as relevant indurations or adverse events. It is suggested to predefine cut-off values above which very large skin test responses are considered systemic adverse events.

There is no evidence that repeated TST increases risk for adverse events.

If two skin tests are given at the same time, then systemic adverse reactions cannot be associated with one test in particular, and relatedness should be ascribed to the new skin test to reduce risk for underestimating harms from the new test.

TST is safe to administer to pregnant women or lactating women. There is no expectation that new skin tests for TB infection cause adverse effects in the fetus or breastfed infant when administered to pregnant or lactating women, and that they can be included when evaluating diagnostic performance of new skin tests for TB infection.

Economic evaluation

New tests should ideally have lower operational costs to the health system and patients compared with existing tests. To evaluate this, costs associated with both the start-up and routine operations of the new test should be considered in studies evaluating the new test. Ideally, an economic evaluation should be incorporated into studies run at demonstration sites or independent research studies. Regardless of the study setting, it is important that an effort is made to document the true costs of implementing the new test as it would be used in practice. The following costs should be considered:

- laboratory equipment and start-up;
- computers and software;
- supplies;
- cold-chain requirements (e.g. reagents, test materials);
- personnel time for different aspects of testing (e.g. obtaining samples, analysis of samples, explaining results to patients);
- initial training (for running the test and interpreting and using the results);
- ongoing training (in-service training to maintain proficiency in test performance);
- quality control and supervision;
- health facility visits by patients with positive or negative tests.

Optimum care models should be considered (e.g. "one-stop shops", where follow-up care for people with positive tests is coordinated and provided on the same day).

Table 2 provides more detail for each of the costs that should be considered for the new test and the reference standard in an economic evaluation.

Table 2. Costs to be considered in economic evaluation of new tests for tuberculosis infection[3]

Type of cost	Reference standard[1]	New test	Additional questions and information
Start-up costs			
Laboratory equipment			If equipment will be used exclusively for testing for TB infection, include total cost of equipment If equipment will not be used exclusively for testing for TB infection, specify approximate proportion of time that it will be dedicated to new test For each piece of equipment, specify how long it is expected to last in order to depreciate capital costs accordingly
Laboratory space			Is additional space required for the new test for TB infection?
Initial calibration of equipment			Specify time requirements and job titles of people involved in task; or give cost of service contract
Licensing			Specify total cost for purchase of licence
Computers			Specify total cost for purchase of all new computers required For each piece of equipment, specify how long it is expected to last in order to depreciate capital costs accordingly
Software			Specify total cost for purchase of all software required
Additional equipment (fridge, air-conditioner, generator)			Specify total cost for purchase of all new equipment For each piece of equipment, specify how long it is expected to last in order to depreciate capital costs accordingly
Initial training			Specify total cost; or number, time requirements and job titles of people who would attend training

3 Not all items are relevant for skin tests and need to be dropped as necessary.

Type of cost	Reference standard[1]	New test	Additional questions and information
Recurring costs			
Ongoing calibration			Specify cost per session and frequency
Ongoing licensing			Specify renewal cost and frequency
Ongoing training			Specify total cost; or number, time requirements and job titles of people who would attend training
Equipment maintenance (per year)			Specify time requirements and job titles of people involved in task; or give cost of service contract
Quality assurance			Specify time requirements and job titles of people involved in task; or give cost of service contract Specify supplies required to conduct quality assurance
Supplies required to administer test (e.g. gloves, syringes)			Ideally specify per-sample cost; or ensure units are otherwise clearly specified
Laboratory supplies for analysis			Ideally specify per-sample cost; or ensure units are otherwise clearly specified
Costs associated with cold chain			If cold chain is required, specify required items and their costs and units
Costs associated with shipping of samples			If samples are shipped to a laboratory, specify shipping costs per sample; specify per-sample costs, or ensure units are otherwise clearly specified
Personnel			
Approximate amount of time required to take sample			Specify time in minutes
Category of personnel who can obtain sample			Specify nurse or other personnel (provide details)
Approximate amount of personnel time required to process and analyse sample			Specify time in minutes
Category of personnel who can process or analyse sample			Specify laboratory technician or other personnel (provide details)
Approximate amount of time to interpret result			Specify time in minutes
Category of personnel who can interpret result			Specify laboratory technician, medical doctor or clerical staff

If a particular cost category is not relevant, indicate as not applicable (N/A).

[1] For the reference standard, start-up costs should be estimated based on the current cost to repurchase equipment, licensing agreements, and so on.

To understand how the new test will be operationalized, the patient flow for an individual who tests positive and for an individual who tests negative should be described in detail. For each scenario, information about how many patient visits are required (e.g. initial meeting, doing the test, providing test results, referral for further evaluation) and how long each takes should be included.

Operational characteristics

Operational challenges in implementing existing tests have been barriers to scale-up of TPT. IGRA needs sophisticated laboratory infrastructure, technical expertise and expensive equipment. TST is considered less resource-intensive than IGRA, but it requires a cold chain, two health-care visits – and providing training for intradermal injection, test reading and interpretation, and quality control is a challenge.

New tests for TB infection should ideally address these operational challenges, particularly if they are not expected to improve predictive performance for development of TB disease. Annex 2 shows the optimal operational characteristics relevant for in vitro tests for TB infection. Ability to deploy at the lowest level of the health-care system is especially important. Instrument-free tests or tests that can be performed with small, portable or handheld instruments that use battery or solar power are needed. Rapid tests would also offer a great advantage.

For skin tests and in vitro tests for TB infection, stability of reagents should be established under different conditions in accordance with the WHO standards for prequalification. It is desirable that reagents can be stored for sufficient periods of time under high temperature and humidity and that a cold chain is not required for transportation. WHO guidance for prequalification of diagnostic assessment is available elsewhere *(20)*.

References

1. Houben RMGJ, Dodd PJ. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. PLoS Med. 2016;1;13(10).
2. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch Intern Med. 2003;163:1009–21.
3. Getahun H, Matteelli A, Chaisson RE, Raviglione M. Latent *Mycobacterium tuberculosis* infection. N Engl J Med. 2015;372(22):2127–35.
4. Implementing the End TB Strategy: the essentials. Geneva: World Health Organization; 2015 (https://www.who.int/tb/publications/2015/The_Essentials_to_End_TB/en/).
5. A/RES/73/3. Political declaration of the High-level Meeting of the General Assembly on the Fight Against Tuberculosis: resolution/adopted by the General Assembly. New York: United Nations General Assembly; 2018 (https://digitallibrary.un.org/record/1649568).
6. WHO consolidated guidelines on tuberculosis: tuberculosis preventive treatment. Geneva: World Health Organization; 2020 (https://www.who.int/publications-detail/who-consolidated-guidelines-on-tuberculosis-module-1-prevention-tuberculosis-preventive-treatment).
7. Rangaka MX, Wilkinson KA, Glynn JR, Ling D, Menzies D, Mwansa-Kambafwile J, et al. Predictive value of interferon-γ release assays for incident active tuberculosis: a systematic review and meta-analysis. Lancet Infect Dis. 2012;12(1):45–55.
8. Matteelli A, Sulis G, Capone S, D'Ambrosio L, Migliori GB, Getahun H. Tuberculosis elimination and the challenge of latent tuberculosis. Presse Med. 2017;46:e13–21.
9. Diel R, Loddenkemper R, Nienhaus A. Predictive value of interferon-γ release assays and tuberculin skin testing for progression from latent TB infection to disease state: a meta-analysis. Chest. 2012;142(1):63–75.
10. Consensus meeting report: development of a target product profile (TPP) and a framework for evaluation for a test for predicting progression from tuberculosis infection to active disease. Geneva: World Health Organization; 2017 (https://www.who.int/tb/publications/2017/tpp_infection_disease/en/).
11. Kik SV, Schumacher S, Cirillo DM, Churchyard G, Boehme C, Goletti D, et al. An evaluation framework for new tests that predict progression from tuberculosis infection to clinical disease. Eur Respir J. 2018;52(4).
12. Use of tuberculosis interferon-gamma release assays (IGRAs) in low- and middle-income countries: policy statement. Geneva: World Health Organization; 2011 (https://www.who.int/tb/publications/tb-igras-statement/en/).
13. Global tuberculosis report 2019. Geneva: World Health Organization; 2019 (https://www.who.int/tb/publications/global_report/en/).
14. Ruhwald M, Aggerbeck H, Gallardo RV, Hoff ST, Villate JI, Borregaard B, et al. Safety and efficacy of the C-Tb skin test to diagnose *Mycobacterium tuberculosis* infection, compared with an interferon γ release assay and the tuberculin skin test: a phase 3, double-blind, randomised, controlled trial. Lancet Respir Med. 2017;5(4):259–68.
15. Barcellini L, Borroni E, Brown J, Brunetti E, Campisi D, Castellotti PF, et al. First evaluation of QuantiFERON-TB Gold Plus performance in contact screening. Eur Respir J. 2016;48:1411–19.
16. Getahun H, Matteelli A, Abubakar I, Abdel Aziz M, Baddeley A, Barreira D, et al. Management of latent *Mycobacterium tuberculosis* infection: WHO guidelines for low tuberculosis burden countries. Eur Respir J. 2015;46:1563–76.
17. Nunn AJ, Phillips PPJ, Meredith SK, Chiang C-Y, Conradie F, Dalai D, et al. A trial of a shorter regimen for rifampin-resistant tuberculosis. N Engl J Med. 2019;380(13):1201–13.
18. Belknap R, Holland D, Feng PJ, Millet JP, Cayla JA, Martinson NA, et al. Self-administered versus directly observed once-weekly isoniazid and rifapentine treatment of latent tuberculosis infection. Ann Intern Med. 2017;167(10):689–97.
19. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010;73:1167–79.
20. Guidance series for WHO prequalification: diagnostic assessment – establishing stability of in vitro diagnostic medical devices. Geneva: World Health Organization; 2017 (https://apps.who.int/iris/bitstream/handle/10665/259742/WHO-BS-2017.2304-eng.pdf?ua=1).

Annex 1

Examples of changes in key assumptions on sample size requirements to demonstrate superiority of tests for sensitivity and specificity

	Reference	New test	Difference	Number required[1] (80% power)
Sensitivity	85%	82%	3%	1164
	85%	80%	5%	430
	85%	77%	8%	174
	80%	77%	3%	1440
	80%	75%	5%	528
	80%	72%	8%	211
Specificity	98%	95%	3%	233
	98%	93%	5%	96
	98%	90%	8%	44
	95%	92%	3%	478
	95%	90%	5%	185
	95%	87%	8%	79

[1] Total number of participants, since both tests are performed in the same person.

Source: Sample size calculator. Vancouver, BC: Department of Statistics, University of British Columbia (https://www.stat.ubc.ca/~rollin/stats/ssize/).

Annex 2

Operational characteristics desirable for new in vitro tests for tuberculosis infection

Number of steps to be performed by operator	< 2; no timed steps
Volume measurements	None
Sample collection and volume	Smallest possible, particularly for children. Pinprick is preferred over phlebotomy. If necessary, a single tube is preferred over multiple tubes
Sample preparation	None or fully integrated
Data analysis	Integrated
Time to results	< 24 hours
Biosafety	Universal precautions
Operating temperature	5–50°C, 90% humidity
Reagents	Self-contained within test kit
Stability of test kit and reagents	24 months at 40°C and 90% humidity; able to tolerate stress during transport (3 days at 50°C)
Instrumentation	Ideally instrument-free test; if not, small, portable or handheld instrument (< 1 kg) that can operate on battery or solar power in places with interrupted power supply
Waste disposal	Standard infected waste disposal at health centre
Internal quality control	Includes positive controls
External quality control	Includes positive and negative controls
Maintenance and calibration	No maintenance or calibration required
Result-capturing, documentation, data display	Ideally instrument-free, but should allow for results to be attached or scanned to reader, saved and printed
Data export (connectivity and interoperability)	Preferably instrument-free, but should allow data export via reader and full data export (on usage of device, error/invalid rates, and personalized, protected results data) over USB port and network Network connectivity through GSM/UMTS mobile broadband modem Results should be encoded using documented standard (e.g. HL7) and formatted as JavaScript Object Notation (JSON) text; JSON data should be transmitted through HTTP(S) to a local or remote server as results are generated Results should be locally stored and queued during network interruptions and sent as a batch when connectivity is restored
Training	1 day dedicated training for non-laboratory-trained health personnel

Adapted from Consensus meeting report: development of a target product profile (TPP) and a framework for evaluation for a test for predicting progression from tuberculosis infection to active disease. Geneva: World Health Organization; 2017.

Stop TB Partnership
New Diagnostics Working Group

World Health Organization

www.who.int/tb
www.stoptb.org/wg/new_diagnostics

9 789240 007185