



Early View

Perspective

Framework for the evaluation of new tests for tuberculosis infection

Yohhei Hamada, Saskia den Boon, Daniela Maria Cirillo, Adam Penn-Nicholson, Morten Ruhwald, Dick Menzies, Olivia Oxlade, Dennis Falzon, Avinash Kanchar, Alexei Korobitsyn, Matteo Zignol, Alberto Matteelli, Tereza Kasaeva

Please cite this article as: Hamada Y, den Boon S, Maria Cirillo D, *et al.* Framework for the evaluation of new tests for tuberculosis infection. *Eur Respir J* 2021; in press (<https://doi.org/10.1183/13993003.04078-2020>).

This manuscript has recently been accepted for publication in the *European Respiratory Journal*. It is published here in its accepted form prior to copyediting and typesetting by our production team. After these production processes are complete and the authors have approved the resulting proofs, the article will move to the latest issue of the ERJ online.

Framework for the evaluation of new tests for tuberculosis infection

Yohhei Hamada ^{1,2}, Saskia den Boon ³, Daniela Maria Cirillo ⁴, Adam Penn-Nicholson ⁵, Morten Ruhwald ⁵, Dick Menzies ⁶, Olivia Oxlade ⁶, Dennis Falzon ³, Avinash Kanchar ³, Alexei Korobitsyn ³, Matteo Zignol ³, Alberto Matteelli ⁷, Tereza Kasaeva³.

¹Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Tokyo, Japan

²University College London, London, United Kingdom

³ Global TB Programme, World Health Organization, Geneva, Switzerland

⁴IRCCS San Raffaele Scientific Institute, Milan, Italy

⁵ Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland

⁶ McGill International TB Centre, Montreal, Canada

⁷ Collaborating Centre for TB/HIV Co-infection and TB Elimination, Dept of Infectious and Tropical Diseases, University of Brescia, Brescia, Italy

Correspondence:

Yohhei Hamada

University College London

Institute for Global Health, 3rd floor, Institute of Child Health, 30 Guilford Street

London WC1N 1EH, United Kingdom

y.hamada@ucl.ac.uk

A 256-character summary

The evaluation of new, improved tests for tuberculosis infection should be expedited using standard study protocols to accelerate adoption into policy and subsequent scale-up. A framework for such evaluation is described.

Abstract

The scale-up of tuberculosis (TB) preventive treatment (TPT) must be accelerated to achieve the targets set by the United Nations High-level Meeting on TB and the End TB Strategy. The scale-up of effective TPT is hampered by concerns about operational challenges to implement the existing tests for TB infection. New simpler tests could facilitate the scale-up of testing for TB infection. We present a framework for evaluation of new immunodiagnostic tests for the detection of TB infection, with an aim to facilitate their standardized evaluation and accelerate adoption into global and national policies and subsequent scale-up. The framework describes the principles to be considered when evaluating new tests for TB infection and provides guidance to manufacturers, researchers, regulators and other users on study designs, populations, reference standards, sample size calculation and data analysis and it is also aligned with the Global Strategy for TB Research and Innovation adopted by the World Health Assembly in 2020. We also briefly describe technical issues that should be considered when evaluating new tests, including the safety for skin tests, costs incurred by patients and the health system patient, and operational characteristics.

Introduction

Treatment of tuberculosis (TB) infection, also known as tuberculosis preventive treatment (TPT), is a critical component to achieve the ambitious targets of the End TB Strategy 2016-2035.[1] Management of TB infection is also critical to pursue TB elimination.[2] Without the prospect that a new, efficacious and safe TB vaccine can be developed and scaled up worldwide in a foreseeable future, continued measures to expand the provision of effective TPT remain of critical importance. Furthermore, 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: 6 million people living with HIV (PLHIV), 4 million children < 5 years who are household contacts of people with TB, and 20 million other household contacts.[3] Progress in expanding TPT coverage among contacts of TB patients has been very limited up to now, with around half a million contacts in 2019.[4] There is no gold standard method for diagnosing TB infection.[5] WHO currently recommends a tuberculin skin test (TST) or an Interferon Gamma Release Assays (IGRA) to test for TB infection.[6] The tests are helpful to identify people at higher risk of developing TB disease and who could benefit from TPT, because in most published studies such risk is higher in people who test positive for TB infection than in those who test negative. [7] However, current tests for TB infection have limited value in predicting the risk of progression from infection to active TB disease.[8]

The development and evaluation of tests characterized by higher prediction capacity (called tests of progression) is a high priority for research. In 2017, WHO published a guidance document on the characteristic of such tests.[9] According to the target product profile defined by WHO, optimal sensitivity and specificity of such tests for predicting development of TB disease are at least 90%,[9, 10] much higher than those of currently available TST or IGRA.[11] Until better tests become available for use under field conditions, existing tests for TB infection, including TST and IGRA, remain the standard tests of choice.

Tests for TB infection are not required before starting TPT in people from high-priority groups such as PLHIV and household contacts aged under five years in high TB burden countries (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 treatment, which carries a risk for adverse events (Figure 1). However, implementation of tests for TB infection is fraught with difficulties, including short supply of quality-assured PPD, the need for training to perform and read TST, and inadequate laboratory set-up to undertake IGRA testing in decentralized settings, and high costs (IGRA). This calls for new tests with better operational characteristics.

New versions of TST and IGRA are already on the market or in the pipeline all using ESAT6 and CFP10 antigens. Diaskintest (Generium, Russian Federation)[12] and ESAT6-CFP10 (Anhui Zhifei Longcom Biopharmaceutical Co. Ltd, China),[13] both of which are IGRA-like skin tests for TB infection, are commercially available and a new IGRA-like skin test, C-Tb (Serum Institute of India, India),[14, 15] has recently been developed. 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 operate with less sophisticated laboratory support.[16]

Evaluation of tests for TB infection is not straightforward due to the lack of a reference standard. Here we present a framework for evaluation of new immunodiagnostic tests for the detection of TB infection, with an aim to facilitate their standardized evaluation and accelerate adoption into global and national policy and subsequent scale-up. The document is primarily intended to guide the work of manufacturers of diagnostics, researchers, funders, regulators, TB programme coordinators, civil society and other stakeholders. The framework was developed by a WHO-convened Technical Expert Group with inputs from reviewers with expertise in this field.[17]

The focus of this framework is on the evaluation of diagnostic performance of tests for TB infection. Additionally, we briefly outline technical issues that should be considered when evaluating new tests for TB infection, evaluation of safety for skin tests, costs incurred by patients and the health system and operational characteristics.

FRAMEWORK FOR EVALUATION OF NEW TESTS FOR TUBERCULOSIS INFECTION

The lack of an adequate gold standard complicates the estimation of sensitivity and specificity, the standard approach to assess the performance of new diagnostic tests. Surrogate reference standards are therefore proposed (see below). A hierarchy of reference standards was developed when WHO reviewed evidence on the use of IGRA (Figure 2).[11]

Reference standards at higher levels of hierarchy should provide stronger evidence for a test's accuracy to identify people who would benefit most from TPT. New tests for TB infection currently in the pipeline are based on similar concepts to TST and IGRA – that is, eliciting an immune response to *M. tuberculosis*-specific antigens either in vivo (size of skin induration) or in vitro (magnitude of cytokine release).[18] They are therefore not expected to offer significant advantage in predicting risk of progression to TB disease. In general, the predictive performance of any new test for TB infection should not be inferior to current technology. Given the challenges to measure predictive value (see also below), studies comparing sensitivity, specificity and concordance between new and current tests may be the best feasible option. In this case, a new test for TB infection could demonstrate non-inferior sensitivity and

specificity based on a pre-specified margin or concordance compared with at least one of the currently available tests endorsed by WHO as rule-in tests for TPT. Because there is mounting evidence that IGRA has higher specificity and possibly higher sensitivity than TST,[19, 20] IGRA should be preferred as a comparator in new trials. [14, 21] Any new tests intended to achieve significant improvement in predictive performance should follow the evaluation framework for tests for predicting progression to TB disease.[9, 10]

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 and could thus inform future policy on the use of new tests (Figure 3). The WHO recommendation to use either TST or IGRA to test for TB infection is based on a review of this type of studies. [6, 22] Using 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. Follow-up of 12 months or longer is recommended to ensure a sufficient length of time for progression to occur. The major ethical issue with this design is that people at risk of TB with a positive test for TB infection should be started on TPT, which interferes with the primary endpoints such as TB incidence or mortality. Although TB incidence can be measured in people who test positive for TB infection but who do not take treatment, this will reduce the study power and may also introduce bias given that people opting out may differ in characteristics associated with endpoints from those who do take treatment. Moreover, the study design may preclude masking of the study subjects (e.g. comparing a new TST to IGRA). In addition, people who do not follow the recommendations from the provider are more likely to discontinue, which may seriously jeopardize the integrity of the cohort design. However, a relative (not absolute) estimate of test predictive performance may still be derived for people opting out of TPT, if two or more tests are evaluated simultaneously in the same cohort. Detailed guidance on how to conduct such studies is available in the framework for evaluation of tests for progression.[9, 10]

Study design 2: sensitivity and specificity using clinical reference standards

Sensitivity can be assessed using culture-confirmed TB as a reference standard. This is important to reduce bias in the estimated accuracy. It should be encouraged even for paediatric TB and forms of extrapulmonary TB that tend to be diagnosed clinically. A diagnosis based only on clinical or

radiographic criteria should only be exceptional. In this case, it is important not to include tests for TB infection in the diagnostic criteria as this will spuriously overestimate the sensitivity of the test for infection.

Specificity depends on the antigens used to induce the immune response. By using overlapping peptides from antigens that are highly *M. tuberculosis*-specific, such as ESAT6 and CFP10, current IGRA tests are, unlike TST, not affected by bacille Calmette–Guérin (BCG) vaccination and their specificity is high at 97-99% based on studies conducted in low-risk populations who are unlikely to have been infected with TB. However, expression of ESAT6 and CFP10 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 the estimation of specificity, the population is 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 example, based on geographical and demographic information. [23] For assessment of tests for TB infection based on TB-specific antigens not found in the BCG vaccine itself (such as ESAT6 or CFP10), vaccination status of participants does not need to be taken into consideration.

The tests should be performed at the same time as the clinical evaluation as a delay between them can introduce bias. [23] 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. As discussed above, estimation of sensitivity and specificity entails two different cohorts, people with active TB and individuals with very low likelihood of TB infection. Participants in each cohort 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 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 concordance.

Study design 3: concordance of tests

This design is essentially a study of agreement between new and reference tests and is the lowest level of evidence for assessment of diagnostic tests. It should be noted that concordance may be low 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 include very young or elderly people, pregnant women, people with more severe disease, or people with serious comorbidities (e.g. HIV, diabetes, renal failure, malnutrition), as these may affect test performance. 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 reference and index tests should be administered in each arm at the same time. Collection of blood or urine specimens for in vitro tests for TB infection should be done at the same time or before administering a TST to avoid a boosting effect.[24]

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. Sensitivity is particularly suboptimal in young children, PLHIV, 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 has excellent specificity, such as the 97-99% specificity of current IGRAs, it would require a large sample size to demonstrate superiority (Table 1).[25] However, a new test could have specificity that is non-inferior to an existing test, and will be preferred because of other advantages. If a non-inferiority design is selected, these other important advantages, such as lower cost to the patient or health system, enhanced feasibility, or point of care availability, must be pre-specified, and measured carefully. It is also important to note that non-inferiority designs should be reserved to test performance when the standard or reference test has very good performance, such as the example of the specificity of the IGRA tests. This is necessary to avoid a risk for the progressive deterioration of test performance over

time if newly emergent technologies are successively compared to the previous generation of test using non-inferiority criteria.[26]

Table1: Example of changes in different key assumptions on sample size requirements to demonstrate *superiority/non-inferiority* of tests for sensitivity and specificity

	Reference	New Test	Difference	Number required* (80% power)
Sensitivity	95%	98%	3%	331
	95%	92%	3%	478
	95%	90%	5%	185
	95%	87%	8%	79
	90%	93%	3%	716
	90%	87%	3%	843
	90%	85%	5%	316
	90%	82%	8%	130
Specificity	98%	99%	1%	1283
	98%	95%	3%	233
	98%	93%	5%	96
	98%	90%	8%	44
	95%	98%	3%	331
	95%	92%	3%	478
	95%	90%	5%	185
	95%	87%	8%	79

From Rollins/STAT/UBC – sample size calculator website

* This the total number of participants - since both tests are performed in the same persons

General determinants of sample size

When designing a study, it is particularly important to pre-specify the effect size, or difference expected, for both superiority and non-inferiority designs. In general, the larger the effect size, the smaller the sample size. It is important to recognize that larger effect sizes may not be realistic and may lead to erroneous conclusions. In a study with too few participants, a new test that has superior performance, but by a small difference from the reference test, may fail to show superiority due to a wide confidence interval crossing the null value. The estimates of performance of the reference test used for sample size calculation should be based on recently conducted high-quality systematic reviews to ensure their accuracy.

Specific sample size calculations and study analysis

Predictive performance

The new test would ideally have good sensitivity and reasonable specificity. 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. While longer periods of follow-up reduce the required samples size, it may result in greater losses to follow-up and 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 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 will have had two or more tests for TB infection, the differences in event rates can be directly calculated, either as a risk difference or as a risk ratio. The incidence rate ratio can be estimated as:

incidence rate among people who test positive/incidence rate among people who test
negative

Sensitivity and specificity

Current tests for TB infection generally have suboptimal sensitivity. Sensitivity is lower in children, PLHIV, and other immunocompromised people. In these populations, superiority of a new test would be preferable to a non-inferiority design.

Current IGRA tests have excellent specificity. Therefore, superior specificity is neither necessary nor likely to be demonstrable. 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.

Sensitivity and specificity can be calculated using the standard formula based on the aforementioned reference standards. However, in a population with very low TB prevalence, specificity can be approximated using the total number of people tested as the denominator instead of the number of people without TB infection as follows:

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

Concordance

If the new test is anticipated to have similar diagnostic accuracy to the reference test but has operational advantages, such as lower 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 especially important when prevalence of positive tests is either very low or very high. 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.[27]

Technical issues

Immunoassays are complex assays influenced by multiple sources of variability which can impact results. Table 1 lists a range of typical sources of variability in in-vitro tests for TB infection like IGRA which should be prioritized by developers in the technical description of the assay. A full list is available elsewhere [17].

Table 2. Typical sources of variability in interferon-gamma release assay-like tests items for documentation.

Sources of variability	Item
Factors impacting stimulation assay	Blood collection tubes: within- and between-lot variability
	Delay in blood processing and incubation time
	Volume of blood
Analytical range of readout assay	LoD
	LLoQ
Imprecision of the readout assay	Intra-assay and inter-assay imprecision, in particular around the cut off for test positivity
Accuracy of readout assay	Recovery
Analytical specificity of readout assay	Cross-reactivity
	Parallelism/dilution linearity

	Common interferents (e.g. rheumatoid factor, lipids, bilirubin, complement, haemolysate)
	Evaluation of curve-fitting model (≥ 5 determinations over multiple runs)
Additional assessments	Inter-laboratory imprecision (reproducibility)
	Analyte stability (Freeze-thaw stability, short-term bench stability, , long-term storage stability)

LOD: Limit of detection; LLOQ: Lower limit of quantification

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 (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 to evaluate changes in IGRA results by using an adequate number of participants without active TB with positive and negative IGRA values representative of normal physiological ranges in cohorts where prior infection is likely and reinfection events are rare. For example, studies in recent adult migrants from high TB incidence countries to low to middle incidence countries where TPT is not routinely initiated for a positive IGRA result could examine reproducibility of the initial IGRA positivity with minimal influence from reinfection. Similarly, reproducibility of IGRA results has been often evaluated in health care workers which demonstrated fluctuation of results.[28] Contacts of people with TB who recently converted to a positive test could also be followed up to assess possibility of reversion. 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 if 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.

The QuantiFERON-TB Gold In-Tube and QuantiFERON-TB Gold Plus assays report considerable variability around the assay cut-off of 0.35 IU/ml.[29] While the cut-off value for any new IGRA should be defined against the concordance with existing tests like QuantiFERON-TB Gold Plus, manufacturers should be cautious about reporting test results around this zone of uncertainty. 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 study zone of uncertainty. 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 populations with very low likelihood of TB infection as well as

people with confirmed TB disease should be used to establish a range of interferon-gamma responses expected for clinically relevant specimens.

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 injection site reactions and other adverse events are similar to or fewer than with TST. Safety should be evaluated in various groups such as PLHIV, children, and pregnant and lactating women. TST is safe to administer to pregnant women and lactating women. Similarly, it is unlikely that new skin tests for TB infection cause adverse effects on fetus or nursing infants when administered to pregnant or lactating women.

The study design should seek to minimize bias in ascertaining local adverse reactions. One method is to give the new and the reference test at the same time in each forearm in a double blinded manner. Another method to assess safety would be to compare adverse events reporting in a randomized controlled trial. However, trials are unlikely to be sufficient to detect rare adverse events and post-marketing surveillance is essential. Methods for recording adverse events should be adequate and clearly recorded e.g. using MedDRA classification (<https://www.meddra.org/>). As skin tests are designed to induce a local reaction it is important to pre-define how to interpret reactions as relevant indurations or adverse events.

Economic evaluation

New tests should ideally have lower health system and patient costs compared to existing tests. In order to evaluate this aspect, costs associated with both the start-up and routine operations of the new test should be considered in studies that are 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” cost of implementing the new test (i.e. 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; personnel time for different aspects of testing; initial and ongoing training; quality control and supervision; and health facility visits by patients with positive or negative tests. Optimum care organization models should be used when implementation is being considered (e.g. ‘one-stop shop’ where follow-up care for those with positive tests is coordinated and provided same day). Additional details on specific costs to consider during evaluations are available presented in the supplementary table.

Operational characteristics

New tests for TB infection should also address operational challenges associated with the existing tests. The 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 a small, portable or hand-held instrument that function by battery or solar power are needed. Rapid tests (e.g. less than one hour for results) would also offer a great advantage. For both skin tests and in vitro tests, the stability of reagents should be established under different conditions in accordance with the WHO standards for prequalification[30]. It is desirable that reagents are stable at high temperature and humidity for sufficient periods and that cold-chain is not required for their transportation. Tests that do not require the withdrawal of venous blood samples would be preferred.

Concluding remarks

The uptake of TPT has been very slow globally partly due to concerns about performance and operational challenges associated with current tests. New tests for TB infection offer an opportunity to facilitate the scale-up and targeting of TPT for more effective use. They also support other initiatives to improve TB preventive treatment, such as the new target product profiles released in 2020. Implementation of this framework for the evaluation of new tests for TB infection will facilitate standardized evaluation of new tests and will expedite their adoption into national policies at the scale needed to achieve global targets.

Acknowledgements

We thank the following individuals for their critical review of the original document published by WHO: Sevim Ahmedov (United States Agency for International Development, USA), Claudia Denkinger (University of Heidelberg, Germany), Darragh Duffy (Institut Pasteur, France), Nazir Ismail (National Institute for Communicable Diseases, South Africa), Afrânio Kritski (Federal University of Rio de Janeiro, Brazil), Elisa Nemes (University of Cape Town, South Africa) and Molebogeng X. Rangaka (University College London, United Kingdom).

References

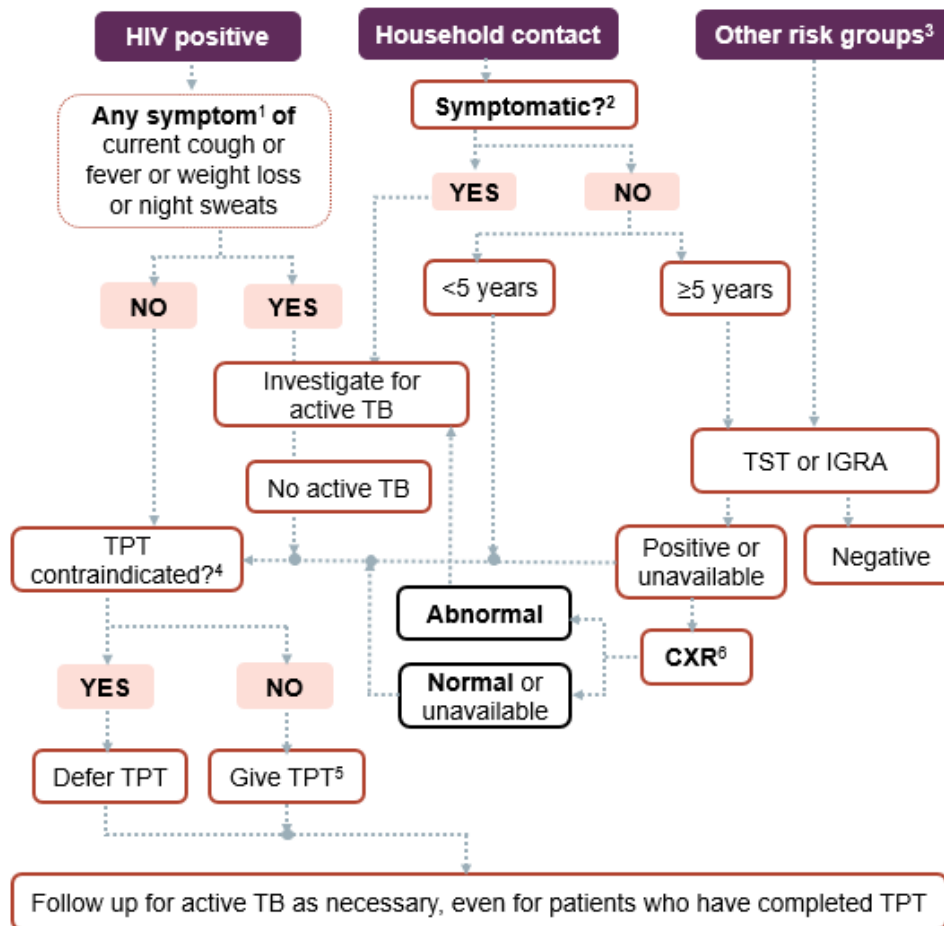
1. World Health Organization. Implementing the End TB Strategy: The Essentials. WHO, Geneva, Switzerland. 2015.
2. Lönnroth K, Migliori GB, Abubakar I, D'Ambrósio L, de Vries G, Diel R, Douglas P, Falzon D, Gaudreau M-A, Goletti D, González Ochoa ER, LoBue P, Matteelli A, Njoo H, Solovic I, Story A, Tayeb T, van der Werf MJ, Weil D, Zellweger J-P, Abdel Aziz M, Al Lawati MRM, Aliberti S, Arrazola de Oñate W, Barreira D, Bhatia V, Blasi F, Bloom A, Bruchfeld J, Castelli F,

- et al. Towards tuberculosis elimination: an action framework for low-incidence countries. 2015; 45: 928–952.
3. 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/164956>). .
 4. World Health Organization. Global Tuberculosis Report 2020. 2020.
 5. Getahun H, Matteelli A, Chaisson RE, Raviglione M. Latent Mycobacterium tuberculosis infection. 2015; 372: 2127–2135.
 6. World Health Organization. WHO consolidated guidelines on tuberculosis: module 1: prevention: tuberculosis preventive treatment. 2020. WHO, Geneva, Switzerland. .
 7. Rangaka MX, Wilkinson KA, Glynn JR, Ling D, Menzies D, Mwansa-Kambafwile J, Fielding K, Wilkinson RJ, Pai M. Predictive value of interferon- γ release assays for incident active tuberculosis: A systematic review and meta-analysis. 2012; 12: 45–55.
 8. Matteelli A, Sulis G, Capone S, D'Ambrosio L, Migliori GB, Getahun H. Tuberculosis elimination and the challenge of latent tuberculosis. 2017. p. e13–e21.
 9. Kik S V., Schumacher S, Cirillo DM, Churchyard G, Boehme C, Goletti D, Rangaka MX, Denkinger CM, Lienhardt C, Gilpin C, Matteelli A, Cobelens F. An evaluation framework for new tests that predict progression from tuberculosis infection to clinical disease. 2018.
 10. World Health Organization. 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. 2017. WHO, Geneva, Switzerland. .
 11. World Health Organization. Use of tuberculosis interferon-gamma release assays (IGRAs) in low- and middle-income countries Policy Statement. 2011. WHO, Geneva, Switzerland. .
 12. Slogotskaya L. Sensitivity of a novel skin test with recombinant protein ESAT6-CFP10 in new cases of tuberculosis in children and adolescents. 2014; 49: S65–S66.
 13. Li F, Xu M, Qin C, Xia L, Xiong Y, Xi X, Fan X, Gu J, Pu J, Wu Q, Lu S, Wang G. Recombinant fusion ESAT6-CFP10 immunogen as a skin test reagent for tuberculosis diagnosis: an open-label,

- randomized, two-centre phase 2a clinical trial. 2016; 22: 889.e9-889.e16.
14. Ruhwald M, Aggerbeck H, Gallardo RV, Hoff ST, Villate JI, Borregaard B, Martinez JA, Kromann I, Penas A, Anibarro LL, de Souza-Galvão ML, Sánchez F, Rodrigo-Pendás JÁ, Noguera-Julian A, Martínez-Lacasa X, Tuñez MV, Fernández VL, Millet JP, Moreno A, Cobos N, Miró JM, Roldan L, Orcau A, Andersen P, Caylá JA. 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. 2017; 5: 259–268.
 15. Aggerbeck H, Ruhwald M, Hoff ST, Borregaard B, Hellstrom E, Malahleha M, Siebert M, Gani M, Seopela V, Diacon A, Lourens M, Andersen P, Dheda K. C-Tb skin test to diagnose Mycobacterium tuberculosis infection in children and HIV-infected adults: A phase 3 trial. 2018; 13: e0204554.
 16. Morten Ruhwald. Future tests for TB infection. Presented in SP-02-Tests for the detection of TB infection - tuberculin skin test vs interferon-gamma release assays. 2020. The 51st Union World Conference on Lung Health. .
 17. World Health Organization. Framework for the evaluation of new tests for tuberculosis infection. 2020.
 18. Hamada Y. State of the art: tests for latent TB infection. AIDS 2020 virtual pre-conference: Turning threats into opportunities: implementing and advancing quality TB services for people living with HIV during the COVID-19 pandemic and beyond. .
 19. Lu P, Chen X, Zhu L, Yang H. Interferon-Gamma Release Assays for the Diagnosis of Tuberculosis: A Systematic Review and Meta-analysis. 2016; 194: 447–458.
 20. Nasiri MJ, Pormohammad A, Goudarzi H, Mardani M, Zamani S, Migliori GB, Sotgiu G. Latent tuberculosis infection in transplant candidates: a systematic review and meta-analysis on TST and IGRA. 2019; 47: 353–361.
 21. Barcellini L, Borroni E, Brown J, Brunetti E, Campisi D, Castellotti PF, Codecasa LR, Cugnata F, Serio C Di, Ferrarese M, Goletti D, Lipman M, Rancoita PM V, Russo G, Tadolini M, Vanino E, Cirillo DM. First evaluation of QuantiFERON-TB Gold Plus performance in contact screening. 2016; 48: 1411–1419.

22. Getahun H, Matteelli A, Abubakar I, Abdel Aziz M, Baddeley A, Barreira D, Den Boon S, Gutierrez SMB, Bruchfeld J, Burhan E, Cavalcante S, Cedillos R, Chaisson R, Chee CBE, Chesire L, Corbett E, Dara M, Denholm J, De Vries G, Falzon D, Ford N, Gale-Rowe M, Gilpin C, Girardi E, Go UY, Govindasamy D, Grant AD, Grzemska M, Harris R, Horsburgh CR, et al. Management of latent Mycobacterium tuberculosis infection: WHO guidelines for low tuberculosis burden countries. 2015. p. 1563–1576.
23. Honda JR, Viridi R, Chan ED. Global Environmental Nontuberculous Mycobacteria and Their Contemporaneous Man-Made and Natural Niches. 2018; 9: 2029.
24. van Zyl-Smit RN, Pai M, Peprah K, Meldau R, Kieck J, Juritz J, Badri M, Zumla A, Sechi LA, Bateman ED, Dheda K. Within-subject variability and boosting of T-cell interferon-gamma responses after tuberculin skin testing. 2009; 180: 49–58.
25. Sotgiu G, Saderi L, Petruccioli E, Aliberti S, Piana A, Petrone L, Goletti D. QuantiFERON TB Gold Plus for the diagnosis of tuberculosis: a systematic review and meta-analysis. 2019; 79: 444–453.
26. Everson-Stewart S, Emerson SS. Bio-creep in non-inferiority clinical trials. 2010; 29: 2769–2780.
27. Watson PF, Petrie A. Method agreement analysis: A review of correct methodology. 2010. p. 1167–1179.
28. Joshi M, Monson TP, Joshi A, Woods GL. IFN- γ Release Assay Conversions and Reversions. Challenges with Serial Testing in U.S. Health Care Workers. 2014; 11: 296–302.
29. Nemes E, Rozot V, Geldenhuys H, Bilek N, Mabwe S, Abrahams D, Makhetha L, Erasmus M, Keyser A, Toefy A, Cloete Y, Ratangee F, Blauenfeldt T, Ruhwald M, Walzl G, Smith B, Loxton AG, Hanekom WA, Andrews JR, Lempicki MD, Ellis R, Ginsberg AM, Hatherill M, Scriba TJ, Rossouw S, Jones C, Schoeman E, Gregg Y, Beyers E, Kruger S, et al. Optimization and Interpretation of Serial QuantiFERON Testing to Measure Acquisition of Mycobacterium tuberculosis Infection. 2017; 196: 638–648.
30. World Health Organization. Guidance Series for WHO Prequalification –Diagnostic Assessment: Establishing stability of in vitro diagnostic medical devices. 2017.

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 a person diagnosed 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 with HIV less than one year of age are only treated for TB infection 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 treatment of TB infection.

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.

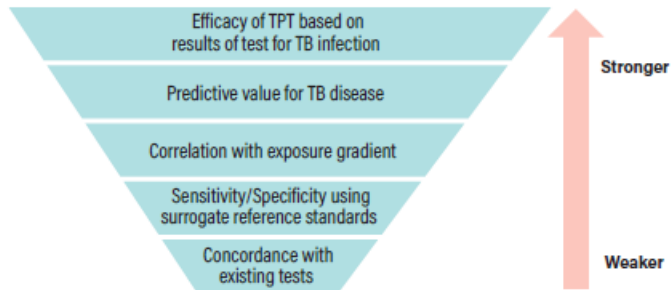
⁴ Including acute or chronic hepatitis; peripheral neuropathy (if isoniazid is used); regular and heavy alcohol consumption. Pregnancy or a previous history of TB are not contraindications.

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: module 1: prevention: tuberculosis preventive treatment. Geneva: World Health Organization; 2020. [5]

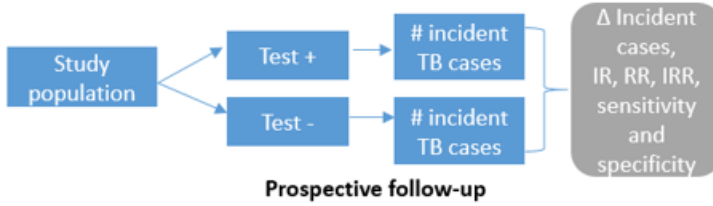
Figure 2. Hierarchy of reference standards used to assess the evidence base for tests for tuberculosis infection



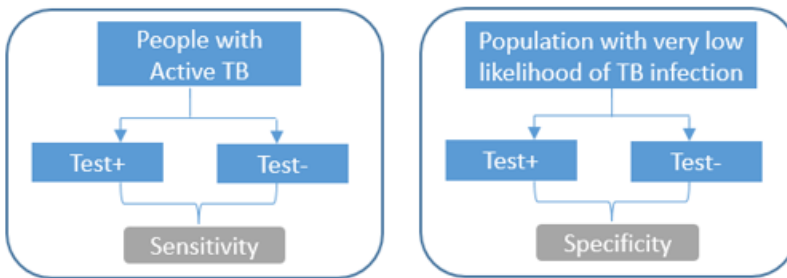
Source: World Health Organization. Framework for the evaluation of new tests for tuberculosis infection. 2020. [12]

Figure 3. Study designs for the evaluation of tests for tuberculosis infection

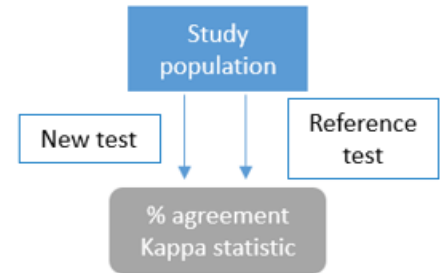
Study 1. Predictive performance



Study 2. Sensitivity and specificity



Study 3. Concordance



Study 1 and 2 include a single test in the figure for simplicity but a new test should be compared with a reference test.

Study design 1 is based on Kik et al, 2018 [9]. See the reference for more details.

RR: risk ratio; IR: incidence rate; IRR: incidence rate ratio.

Supplementary table. Costs to be considered in economic evaluation of new tuberculosis infection tests

Type of cost	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
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

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