Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Guidelines

Tuberculosis treatment monitoring tests during routine practice: study design guidance

Emily Lai-Ho MacLean [1], Alexandra J. Zimmer [2], Saskia den Boon [3], Ankur Gupta-Wright [4], Daniela M. Cirillo [5], Frank Cobelens [6], Stephen H. Gillespie [7], Payam Nahid [8], Patrick P. Phillips [8], Morten Ruhwald [9], Claudia M. Denkinger [10], [11], *

[1] *NHMRC Clinical Trials Centre, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia*
[2] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada*
[3] *Global Tuberculosis Programme, World Health Organization, Geneva, Switzerland*
[4] *Institute for Global Health, University College London, London, UK*
[5] *Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy*
[6] *Department of Global Health and Amsterdam Institute for Global Health and Development, Amsterdam University Medical Centers Location, University of Amsterdam, Amsterdam, The Netherlands*
[7] *Division of Infection and Global Health, School of Medicine, University of St Andrews, St Andrews, UK*
[8] *Center for Tuberculosis, University of California San Francisco, San Francisco, CA, USA*
[9] *FIND, Geneva, Switzerland*
[10] *Division of Clinical Tropical Medicine, Center of Infectious Diseases, Heidelberg University Hospital, Heidelberg, Germany*
[11] *Center of Infection Research (DZIF), Partners Site Heidelberg, Heidelberg, Germany*

ARTICLE INFO

Article history:
Received 15 September 2023
Received in revised form
10 December 2023
Accepted 25 December 2023
Available online 3 January 2024

Editor: E.J. Kuijper

Keywords:
Recommendations
Reporting
Study design
Testing
Treatment monitoring
Tuberculosis

ABSTRACT

Scope: The current tools for tuberculosis (TB) treatment monitoring, smear microscopy and culture, cannot accurately predict poor treatment outcomes. Research into new TB treatment monitoring tools (TMTs) is growing, but data are unreliable. In this article, we aim to provide guidance for studies investigating and evaluating TB TMT for use during routine clinical care. Here, a TB TMT would guide treatment during the course of therapy, rather than testing for a cure at the regimen's end. This article does not cover the use of TB TMTs as surrogate endpoints in the clinical trial context.
Methods: Guidelines were initially informed by experiences during a systematic review of TB TMTs. Subsequently, a small content expert group was consulted for feedback on initial recommendations. After revision, feedback from substantive experts across sectors was sought.
Questions addressed by the guideline and Recommendations: The proposed considerations and recommendations for studies evaluating TB TMTs for use during the treatment in routine clinical care fall into eight domains. We provide specific recommendations regarding study design and recruitment, outcome definitions, reference standards, participant follow-up, clinical setting, study population, treatment regimen reporting, and index tests and data presentation. Overall, TB TMTs should be evaluated in a manner similar to diagnostic tests, but TB TMT accuracy must be assessed at multiple timepoints throughout the treatment course, and TB TMTs should be evaluated in study populations who have already received a diagnosis of TB. Study design and outcome definitions must be aligned with the developmental phase of the TB TMT under evaluation. There is no reference standard for TB treatment response, so different reference standards and comparator tests have been proposed, the selection of which will vary depending on the developmental phase of the TMT under assessment. The use of comparator tests can assist in generating evidence. Clarity is required when reporting of timepoints, TMT read-outs, and analysis results.
Implementing these recommendations will lead to higher quality TB TMT studies that will allow data to be meaningfully compared, thereby facilitating the development of novel tools to guide individual TB therapy and improve treatment outcomes. **Emily Lai-Ho MacLean, Clin Microbiol Infect 2024;30:481**

* Corresponding author. Claudia M. Denkinger, Division of Clinical Tropical Medicine, Center of Infectious Diseases, Heidelberg University Hospital, Im Neuenheimer Feld 672, 69120, Heidelberg, Germany.
E-mail address: Claudia.Denkinger@Uni-Heidelberg.de (C.M. Denkinger).

https://doi.org/10.1016/j.cmi.2023.12.027

Scope

Effective tuberculosis (TB) regimens have been available for over 50 years, but treatment outcomes are still suboptimal. Treating TB, especially drug-resistant TB (DR-TB), is challenged by complex and toxic regimens [1]. Treatment failure occurs because of a variety of factors that, aside from causing individual-level morbidity and mortality, heighten the risk of disease relapse, transmission, and development of DR-TB [2,3]. The WHO reports that treatment success varies from 74% to 91% among people starting treatment for drug-susceptible TB (DS-TB), and 56% to 69% among those starting treatment for DR-TB [4]. There is, therefore, a need to improve treatment outcomes, and people undergoing routine TB treatment may benefit from the use of TB treatment monitoring tools (TMTs) that could guide the treatment they receive.

However, there is no accurate TB TMT available for patients in routine clinical care, although research in this area is growing [5,6]. In this document, we aim to provide recommendations regarding the design and reporting of studies investigating TB TMTs in the clinical setting, where patient treatment would be guided by the results of the TB TMT. This is in contrast to the clinical trial setting, where a TB TMT may be used to assess a surrogate endpoint, or for other uses. These guidelines are applicable to studies evaluating the use case of a TB TMT to direct treatment during the course of therapy, rather than the use case of assessing disease severity at treatment initiation or testing for cure at the regimen's completion.

Context

A test or tool capable of assessing an individual's response to anti-TB therapy (Table 1) during routine care would inform healthcare workers whether a patient is responding favourably or unfavourably to TB treatment, allowing for a change in treatment with the aim to improve treatment outcomes, shorten treatment, and/or reduce treatment-related morbidity and the development of drug resistance. Ideally, a read-out of this tool measured once early in treatment would accurately predict the final treatment outcome, but with current technology, multiple readings throughout treatment may be needed [7]. Indeed, such repeatedly measured markers are referred to by the United States Food and Drug Administration as ‘monitoring biomarkers’ and can be used to ‘assess response of a disease […] to a treatment’ [8]. Biomarkers, multi-marker biosignatures, clinical features, and imaging outputs are some possible measurands (alone or in combination comprising a score) that could be included in a TMT, although measurands are not limited to these possibilities. Previous reviews have summarized the pipeline of TB TMTs and strategies [5,9–11]. Optimally, these measurands could also be used in culture-negative patients, children, or when a suitable sample for culture cannot be obtained.

In 2021, WHO updated the TB treatment outcome definitions [12]. Outcomes are assessed using culture with or without smear microscopy for DR-TB and DS-TB during therapy and at treatment end. However, obtaining results from culture takes weeks, at which time the treated patient may be difficult to re-contact [13], and smear microscopy cannot differentiate viable from non-viable bacteria [14]. In addition, both techniques rely on sputum as a sample, which is difficult to collect from individuals who have responded well to treatment, people living with HIV [15], and children [16], and is uninformative for patients with extra-pulmonary TB without concomitant PTB [17]. Because of these tools' significant limitations, both techniques are under-implemented as TB TMTs [18–20].

There has been growing research into biomarkers that correlate with response to TB treatment, as documented in a recent systematic review [6]. The review may be consulted for biomarkers and assays that have already been evaluated. However, there were substantial differences in reporting, and the risk of bias was high in most publications, impeding the direct comparison of findings between studies. In part, this can be attributed to a lack of available guidance for the development and evaluation of TB TMTs. Target product profiles (TPPs) have previously been published to guide the development of TB tests for certain use cases (e.g. a non-sputum-based biomarker diagnostic test [21] or a peripheral drug susceptibility test [22]), but only in 2023 were TPPs for routine treatment monitoring tests developed and published [7]. Available guidance for evaluating TMTs has focused on the clinical trial use case: (a) a policy document for tests measuring bacteriological response in the clinical trial context from a 2011 United States Centers for Disease Control and National Institutes of Health meeting [23] and (b) a TPP for a treatment monitoring test to be used in clinical trials from Unite4TB (article in press).

Question addressed by the guideline

Here, we aim to provide guidance for studies investigating and evaluating TB TMTs that could be used during routine clinical care in

Table 1
Terminology and definitions used in this manuscript

Term	Working definition
Treatment monitoring tool, treatment monitoring test[a]	Tests or tools of TB treatment optimization during TB treatment that aim to identify patients at high risk of poor outcomes because of inadequate response to their current TB treatment, and whose outcome would be improved with differentiated or optimized TB treatment[b]. A treatment monitoring tool may utilize a single or multiple measurand(s), and may or may not require biological specimens.
Discovery phase study	Studies that aim to determine the presence of a signal to be used as a novel treatment monitoring tool.
Early-phase study	Studies that aim to determine whether a novel treatment monitoring tool can predict patient treatment outcome by measurement early in the course of therapy. Results from a TMT at this stage of evaluation would not be used to inform an individual's treatment.
Late-phase study	Studies that aim to determine whether a novel treatment monitoring tool can improve patient management and lead to improved treatment outcomes, as the tool's result will guide treatment administered to individuals in the study.
Bacteriologic conversion[c]	This describes a situation in a patient with bacteriologically confirmed TB where at least two consecutive cultures (for DR-TB and DS-TB) or smears (for DS-TB only), taken on different occasions at least 7 days apart, are negative.
Treatment failure[c]	A patient whose treatment regimen needed to be terminated or permanently changed (because of no clinical response and/or no bacteriological response, adverse drug reactions, or evidence of additional drug resistance to medicines in the regimen) to a new regimen or treatment strategy.
Relapse[d]	Relapsed patients with TB have previously been treated for TB, were declared cured or treatment completed at the end of their most recent course of treatment, and are now diagnosed with a recurrent episode of TB (either a true relapse or a new episode of TB caused by reinfection).
Early relapse	As above for ‘relapse’, but restricted to cases arising up to 6 mo after therapy completion to aid in distinguishing true relapses from reinfections.

DR, drug-resistant; DS, drug susceptible; TB, tuberculosis; TMT, treatment monitoring tool.

[a] Use cases and target characteristics for these tests/tools are described in further details in the WHO Target Product Profiles for Tests for Tuberculosis Treatment Monitoring and Optimisation [7].

[b] TB treatment includes adjunctive therapy or change in general management, e.g. more intensive follow-up.

[c] WHO definitions available in [10].

[d] WHO definition available in [11].

high TB burden settings. These tools are intended to help healthcare workers assess the effectiveness of TB treatment and, consequentially, improve treatment outcomes. Following the corresponding TPP [7], this guidance is applicable at ‘minimum’ to studies evaluating TB TMTs among a target population of people with pulmonary TB or any form of bacteriologically confirmed TB, but ‘optimally’ to all people already receiving treatment for TB. Recommendations will concern study design and reporting elements.

Methods

The current project was initiated by the Biomarkers Task Force of the New Diagnostics Working Group (NDWG). NDWG is a multisector group that aims to progress TB diagnostics; NDWG receives funding from the Stop TB Partnership and United States Agency for International Development (USAID). Initial drafting of recommendations was conducted by a core group of multidisciplinary experts from McGill University, University of Sydney, and Heidelberg University who had recently performed several rounds of critical review of the TB treatment monitoring literature during the process of conducting a systematic review of biomarkers for TB treatment monitoring [6]. The experiences and observations made while systematically reviewing the literature and extracting data from eligible publications served as the basis of evidence for the initial draft guidance manuscript. All details regarding the search strategy, methods, risk of bias assessment, and results are available in the published systematic review [6]. Subsequently, the initial draft was shared for several rounds of feedback with substantive experts from FIND, a non-profit product development partnership focused on diagnostics, University of California San Francisco, IRCCS San Raffaele Scientific Institute, WHO, and University College London, all of whom are included in the document's authors list. In a final step, recommendations from the revised, late-stage draft were presented at a meeting in September 2022 to attendees assembled by WHO for the development of the TPPs for tests for TB treatment monitoring and optimization [7]. Members represented government, industry, non-governmental agencies, product development partnerships, research institutes, and universities. A full list of meeting participants is available in the published TPPs [7].

Recommendations for studies evaluating TB TMTs for routine care

The selected TMT would be tightly correlated with an individual's response to treatment, so that a read-out of the tool's measurand accurately predicts treatment outcome. Evaluation of a TMT can be approached in a similar, but not identical, manner to the evaluation of a novel diagnostic test, that is, the accuracy of the tool can be ascertained by comparing it with a reference standard; at present, this is compromised by the poor predictive value of the available tools. Although evaluating a diagnostic test requires a measurement taken at a single timepoint, candidate TMTs should be evaluated throughout the course of treatment to understand how well they reflect an individual's response to therapy. In addition, although the population selected for studies evaluating diagnostic tests should comprise individuals undergoing evaluation for TB, studies evaluating novel TMTs should target individuals who are already undergoing treatment for TB.

Recommendations for study design and recruitment

Evaluating a new TMT comprises multiple phases. In ‘discovery’ studies and ‘early-phase’ studies (Table 1), a biomarker or other measurand may be identified as having the potential to predict treatment outcome. Once a candidate TMT is found to have promising accuracy in predicting treatment outcome, later phase studies will be necessary to determine the TMT's usefulness in making treatment decisions. Across study phases, participant recruitment timing and sampling strategy should always be clearly described. Convenience sampling should be avoided to limit selection bias [24].

In discovery phase work, where a novel signal is being identified, retrospective studies of well-characterized bio-banked specimens (e.g. from a TB treatment trial) would play an important role, because using existing specimens and sampling based on known treatment outcome (i.e. as is the practice of case-control studies) can expedite assessment. In general, retrospective studies of well-characterized bio-banked specimens will likely be a necessary starting point for identifying novel signals for TMTs. However, beyond the discovery stage, specimens should not be selected based on treatment outcome because of the high risk of spectrum bias [25], which may lead to over-estimates of sensitivity and specificity. To minimize potential spectrum bias, ‘control’ samples from patients with a range of disease severity (e.g. as demonstrated by chest radiography, smear status, clinical severity, or other measures) should be selected, to reflect differing extents of TB disease.

Once a signal is characterized, the TMT's accuracy in predicting treatment outcome should be verified in early verification phase studies prospectively with fresh samples in a single-gated manner, i.e. all participants/specimens enter the study based on one common set of inclusion criteria (e.g. patients with culture-positive TB) [26].

Subsequently, promising TMTs should be verified in early validation clinical studies utilizing consecutive recruitment in patient populations of intended use. For example, a study validating a putative TMT could recruit participants from a cohort where all participants' treatment outcomes will eventually be observed (e.g. a treatment trial of novel TB treatment regimens; a pragmatic study with more intensive follow-up of people receiving treatment through their country's national TB program). The TMT's results could then be compared with the treatment outcomes, with sensitivity and specificity for predicting treatment outcome computed at various timepoints. In these prospective early-phase validation studies, multiple candidate TMTs could be evaluated and compared head-to-head. During this phase, physical elements, technical elements, and standard operating protocols of the TMT should be locked. This phase would be followed by a review from independent experts and, if recommended, could undergo later phase validation. Critically, at this early validation phase, the TMT's output would not be used for patient management.

In later phase validation studies, the aim is to determine whether deploying the TMT to inform treatment, as part of a clinical decision-making algorithm, alone or in combination with other interventions (e.g. enhanced clinical monitoring), will improve patient outcomes (see section ‘Recommendations for a reference standard to assess treatment response’ for details), compared with administering treatment following standard-of-care procedures. The most definitive way to assess if a TMT improves patient outcomes is with prospective randomized studies. For example, peripheral clinics with similar patient profiles could be cluster-randomized to use or not use a TMT to guide the management of newly diagnosed patients with TB. Rates of treatment failure and/or early relapse in each arm could then be compared. Alternative study designs (e.g. adaptive) utilizing different randomization strategies (e.g. individual randomization) are possible and should be considered. Superiority and non-inferiority trial designs should be considered and adopted as most appropriate, given the TMT and specific research question at hand (see section ‘Recommendations for a reference standard to assess treatment response’ for discussion of endpoints). Other considerations, such as cost or treatment burden to patient, may also be relevant.

Recommendations for a reference standard to assess treatment response

Studies must clearly define treatment response or outcome. For TMT evaluation, we suggest using different definitions of treatment response as the reference standard depending on the TMT's evaluation phase.

In discovery phase studies ascertaining whether a biomarker can accurately predict a patient's eventual treatment outcome, ‘bacteriological conversion’ [10] (Table 1) is the best assessment of response. Biomarkers displaying high agreement with bacteriological conversion could advance for further evaluation as a TMT.

In both early and later validation phase studies, we recommend using a composite measure of the treatment outcomes ‘treatment failure’, ‘early relapse’ (Table 1), and death (with reporting stratified by TB-related reason or not) as a reference standard. Participants would thus be classified as experiencing this composite poor outcome, or not. Note that WHO's definition of ‘treatment failure’ is itself a composite outcome that encompasses multiple reasons why a patient may not be cured. For research purposes, patient outcomes must be reported with sufficient detail so that readers can understand whether treatment failure resulted from lack of clinical and/or bacteriological response, or other reasons, e.g. therapy cessation or modification because of adverse drug reactions or drug resistance. We recommend using the WHO definition of ‘relapse’ after 6 months post-treatment to distinguish likely relapse from reinfection. Limiting follow-up time to 6 months is a pragmatic decision as the likelihood of reinfection increases over time [27]. Optimally, however, genotyping or sequencing methods would provide a superior distinction between reinfection and relapse, and we recommend sequencing for all observed recurrent cases [28].

We advise against using ‘treatment success’ (a composite outcome of ‘cured’ and ‘treatment completed’) or its components as an outcome. ‘Cure’ requires ‘evidence of bacteriological response’, which is often unobservable in many people with TB, whereas ‘treatment completed’ refers to adherence to national guidelines, lacking direct clinical relevance to patient health.

All study participants should have their TMT results compared with the same reference standard to avoid partial verification bias [29]. If different participants received different reference standards, this must be reported. When ascertaining an individual's treatment outcome, those individuals independently reviewing outcomes, e.g. expert clinical panels, should be blinded to the results of the TMT under evaluation. Blinding should be described in the study methods.

Recommendations for comparator tests

Evaluating the TMT's concordance with one or multiple comparator tests is valuable evidence, particularly as there is no perfect test for measuring TB treatment response, although it remains less convincing than comparison with treatment outcome (early studies) or impact on treatment outcome (late studies). Comparator tests permit ‘benchmarking against a test with the same intended use for which a large evidence base exists’ [30], aiding in the interpretation of a new tool's performance. Although culture could be used as a reference standard for discovery phase work, in later phase studies, on its own it should be considered as more of a comparator and not a perfect reference standard. This is because culture can only predict bacteriological response among sputum-producing people, and even in this subgroup it cannot perfectly predict treatment outcome [14,31]. Culture combined with other comparator tests and clinical improvement (e.g. C-reactive protein) would provide stronger evidence of treatment response than culture alone. Alongside culture, radiography or computer tomography could provide insight into the clinical response [32], whereas culture would assess the bacteriological response and sterilizing cure. Data on position-emission tomography/computer tomography highlight that clinical cure might not be associated with complete bacterial clearance in certain cases [33,34]. PCR methods with a DNA target are generally unable to differentiate between live and dead bacteria, and intercalating dyes such as propidium monoazide have shown limited benefit [35]. Although imperfect, Xpert (either MTB/RIF or MTB/RIF Ultra) (Cepheid, Sunnyvale, California, USA) could be used to supplement culture results, because limited data exist on the use of semi-quantitative cycle threshold-values from Xpert (MTB/RIF or Ultra) to measure change in bacterial burden from baseline, but its performance is imperfect [36–38]. Although no published data exist, it is conceivable that the use of cycle threshold-values could be feasible using other WHO-recommended molecular platforms. Recently, TB molecular bacterial load assay, a method that detects *Mycobacterium tuberculosis* 16S ribosomal rRNA, has been shown to correlate closely with the Mycobacterial growth indicator tube (MGIT) liquid culture time to positivity [39] and early relapse [40]. The rRNA synthesis (RS) ratio, which measures ongoing rRNA synthesis, has been shown to reflect the rate of *M. tuberculosis* replication [41] and it is now being explored for use in informing TB treatment shortening [42]. Finally, as smear microscopy is frequently performed in programmatic settings, its inclusion as a comparator in late-stage studies may be useful.

Recommendations for patient follow-up frequency and timing

The lack of standardized timepoints at which samples are collected during patient follow-up impedes the comparison of TB treatment monitoring studies [6]. Thus, a more consistent approach would be favourable.

Putative TMTs should be systematically evaluated at different timepoints to identify how the tool's measurand changes throughout treatment and when the TMT would have the highest accuracy. This will necessitate collecting specimens at multiple timepoints throughout treatment. Selecting timepoints will be influenced by the TMT's measurand, because certain biomarkers might provide a signal indicating that an individual is responding appropriately to treatment within the first 2 weeks of treatment [39,43,44], whereas other measurands may take longer. Some TMTs may require a single reading, whereas others may require a baseline measurement for comparison. The cost of repeat readings should be balanced with the ease of implementation and patient impact, in line with the TPPs for a TMT [7]. Table 2 shows suggested follow-up timepoints for when a candidate TMT should be evaluated in development. Ultimately, the ideal timing and frequency to use a TMT will be tool specific and will be determined by data from well-designed studies, which may not align with the timepoints shown in Table 2.

Recommendations for study location (clinical setting)

Different TMTs will be applicable to different clinical settings, although DS-TB is typically monitored in more peripheral clinical settings or in the community. With the increasing simplification of DR-TB therapy [45–47], DR-TB TMTs may become deployable in more peripheral settings. Studies evaluating novel TMTs should describe the setting in which the treatment monitoring test was performed. This includes a description of the geographic location, as well as the clinical setting (e.g. secondary or tertiary hospital, peripheral clinic) and training level of test operators. If the index test was performed at a different site from where the TB samples were collected, this must be mentioned. In later phase studies, a multi-site design will help understand if the tool may be deployed

Table 2
Recommended follow-up timepoints for candidate treatment monitoring tool evaluation

	Timepoints at which candidate TMT should be evaluated during development—minimal	Timepoints at which candidate TMT should be evaluated during development—optimal
Timing	• At treatment initiation • 4 wk • 8 wk • EOT • 3 mo post EOT • If applicable, at the time of TB recurrence	• At treatment initiation • 2 wk • 4 wk • 8 wk • Halfway through treatment (if different from 8 wk) • EOT • 3 mo post EOT • 6 mo post EOT • 12 mo post EOT • If applicable, at the time of TB recurrence

EOT, end of treatment; TB, tuberculosis; TMT, treatment monitoring tool.

across different geographic settings, particularly because treatment success rates may vary by region. High TB burden settings should be prioritized to enable generalizability to other settings where the TMT may provide the most benefit.

Recommendations for study population

Because TB TMTs are intended to direct an individual's patient's treatment, studies evaluating TB TMTs should include people who have received a diagnosis of TB and are already undergoing treatment. Studies evaluating TB TMTs should avoid comparing a putative TMT's performance between people with TB and people without TB, because this is not the relevant comparison for a test's eventual deployment.

Some biomarkers may present differently across different patient populations, e.g. people living with HIV [48], children [49], people with recurrent TB [50], or between strains [51]. Thus, it is important to collect key demographic and clinical characteristics that may be relevant to the biomarker(s) under investigation for all phases of evaluation. HIV status, baseline drug resistance profile, history of TB disease, and regimen should always be reported, given their relevance to TB disease progression and severity. Other characteristics, such as diabetes, nutrition status, and Bacillus Calmette–Guérin status, should be reported when present in the study population.

Inappropriate exclusion of participants should be prevented, e.g. based on smear status, to avoid selection bias. If the study excludes particular populations, such as smear-negative individuals, this must be clearly stated to understand the generalisability of findings.

Participant selection should reflect the phase of the TMT's investigation. Discovery phase studies may select bio-banked samples from people whose TB treatment outcomes were ‘cured’ and, as a control group, ‘failed’ [12]. Contrastingly, a late-phase clinical utility study should consecutively enrol participants who will comprise the eventual target population, e.g. all adults treated for pulmonary or extra-pulmonary TB (Table 3).

Recommendations for treatment regimen and treatment adherence programmes

Existing TB treatment regimens for DS-TB [52] and DR-TB [53] vary in duration, dose, and combination of antibiotics. At a minimum, studies should identify the different regimens used and the number of patients on each treatment regimen. In addition, treatment adherence as well as the use of tools that promote treatment adherence (with description and reference) should be reported. Sensitivity analyses, including individuals with low treatment adherence rates, may produce estimates that reflect ‘real-world’ conditions. In late-stage studies, adequate description of actions taken based on TMT results must be provided.

Recommendations for index test and data analysis

The TMT result should be reported at each follow-up point for all study participants. When a commercial assay is being evaluated as a TMT (e.g. a commercial CRP assay), the assay's product name, manufacturer, and version/lot number should be provided. For research-use-only assays or assays that are not design-locked, sufficient procedural information should be provided. In later phase studies, the tool under evaluation should be design-locked. The cut-off (including measurand units) used to indicate treatment response at each timepoint must be reported. Some commercial assays have a pre-specified threshold for detecting TB positivity (e.g. grade 1 for a lateral flow urine lipoarabinomannan (LAM) assay [54]), which are used at initial presentation. It must be specified if this threshold is also used for treatment monitoring purposes. If a TMT's output is quantitative, it should be reported in addition to the cut-off used (i.e. not just binary positive/negative).

If no cut-off is predefined (as can be expected in discovery and early-phase studies), the change in measurand over time should be reported using measures of central tendency (mean or median) and spread (standard deviation or interquartile range). These values should be reported in a table with units. Graphs that display this

Table 3
Summary of recommendations for studies evaluating tools for TB treatment monitoring

Domain	Key recommendation(s)
1. Study design and recruitment	The development phase of the TMT (i.e. discovery, early, late) must guide the design of the evaluation study and strategy for participant selection. E.g., case-control studies using bio-banked samples are important for discovery phase work; prospective studies using consecutive sampling would have utility for early-phase studies; randomized designs can be used for late-phase studies.
2. Reference standard to assess treatment response	A study's reference standard will depend on the development phase (i.e. discovery, early, late) of the TMT under evaluation, but WHO treatment outcomes should be used as reference standards.
3. Comparator tests	Comparator tests such as smear microscopy, culture, and imaging can help benchmark the performance of new TB TMTs.
4. Patient follow-up frequency and timing	Suggested timepoints to assess new TB TMTs: treatment initiation, 4 wk, 8 wk, end of treatment, 3 mo post end of treatment, at the time of TB recurrence (if applicable).
5. Study location (clinical setting)	Clearly describe the clinical setting and healthcare level of the TMT study.
6. Study population	People diagnosed with TB who are taking TB treatment should be included in studies of TB TMTs.
7. Treatment regimen and treatment adherence programmes	Clearly report study participant treatment regimens and rates of treatment adherence, as well as any adherence support provided to study population members.
8. Index test and data analysis	Present TB TMTs performance at each follow-up timepoint in 2-by-2 contingency tables and as accuracy estimates.

TB, tuberculosis; TMT, treatment monitoring tool.

change over time may be used to visualize the change in addition to a table.

Further stratification at each follow-up is encouraged for the following variables: HIV status (with CD4 count), baseline presence drug resistance, baseline smear status, history of TB disease, TB regimen used, level of treatment adherence, and sex.

The accuracy of TMTs for a given treatment outcome (see section 'Recommendations for a reference standard to assess treatment response') should be evaluated at all follow-up times using (a) a previously identified analytical cut-off from the literature and/or (b) the best-performing threshold (e.g. Youden's Index) while aiming to achieve the targets in the WHO TPPs for tests for TB treatment monitoring and optimization [7]:

- A threshold that provides ≥90% sensitivity optimally, or ≥75% sensitivity minimally; and
- A threshold that provides ≥90% specificity optimally, or at ≥80% specificity minimally.

For all phases of studies, two-by-two tables and/or receiver operating characteristic curves with areas under the curve should be presented at each follow-up timepoint to assess the accuracy versus the chosen reference standard. Appropriate within-subject analyses may be necessary to mitigate the increased risk of type I errors, e.g. repeated measures analysis of variance [55]. Although a 'global' area under the curve avoids multiplicity issues, it obscures the TMT's performance at each follow-up point. When reporting on TMT accuracy, we recommend following Standards for Reporting Diagnostic accuracy studies (STARD) guidelines to the extent applicable [56].

Closing remarks

Future studies investigating novel biomarkers and tools for routine TB treatment monitoring should incorporate the above design and reporting recommendations. A summary of key recommendations is provided in Table 3. This will enhance data utility by reducing bias and improving data quality. TMTs should be evaluated based on their ability to improve patient treatment outcomes. This will enable assessments using the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) framework as commonly used in WHO guideline reviews [57].

Although new tools for treatment monitoring may increase rates of successful treatment, we acknowledge that test results are only one element within a larger constellation of considerations clinicians must consider, which include factors beyond this work's scope. We have attempted to propose these guidelines in a purposefully general manner in an attempt to provide useful recommendations that would be applicable to studies evaluating any possible format of TB TMT, regardless of study population subgroups, healthcare level, or routine clinical algorithms.

Some uncertainty persists in the conduct of treatment monitoring studies. This is largely because of the large possible variety in formats of potential TMTs and the limitations of the reference standard. Our recommendations strive to provide optimal and standardized guidance considering these realities. This will allow for a more standardized approach to evaluating tools for treatment monitoring, which in turn will help streamline the development of new and accurate TMTs.

Author contributions

EL-HM, AJZ, and CMD conceptualized the manuscript and drafted the initial set of recommendations. EL-HM, AJZ, SdB, and CMD wrote the original draft manuscript. SdB, AG-W, DMC, FC, SHG, PN, PPP, and MR commented and provided feedback on multiple drafts of the manuscript. CMD supervised the project. Funding was acquired by EL-HM, DMC, MR, CDM. EL-HM presented the advanced draft recommendations at a Technical Consultation for the Development of Target Product Profiles for Tests and Biomarkers for Tuberculosis Treatment Monitoring and Optimisation in September 2022; the list of meeting attendees of the meeting is available in [7]. Funding was acquired by EL-HM, DMC, MR, and CMD. The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliate.

Transparency declaration

CMD reports project-specific funding from WHO; grants for various projects on TB diagnostics development and evaluation support from FIND, Geneva; grants for Rapid Research in Diagnostics Development (R2D2) for TB network from National Institutes of Health (NIH) US; grants for various projects on TB diagnostics development and evaluation from German Center for Infectious Disease Research (DZIF); being a Member of the WHO Advisory Group on Tuberculosis Diagnostics and Laboratory; being Vice Chair in Heidelberg and TB Co-Chair of DZIF; and being Academic Editor of PLOS Medicine. EL-HM reports support for this project from the New Diagnostics Working Group—Biomarkers Taskforce. SHG reports that a patient is in preparation for some components of an antibiotic treatment monitoring tool at the UK Patent office. Other authors declare no conflicts of interest. Funding for the study was provided by the New Diagnostics Working Group—Biomarkers Taskforce. The New Diagnostics Working Group was supported by funding received from the Stop TB Partnership and USAID.

Acknowledgements

We are grateful to the participants attending the Technical Consultation for the Development of Target Product Profiles for Tests and Biomarkers for Tuberculosis Treatment Monitoring and Optimisation [7] for their feedback.

References

[1] Forget EJ, Menzies D. Adverse reactions to first-line antituberculosis drugs. Expert Opin Drug Saf 2006;5:231–49. https://doi.org/10.1517/14740338.5.2.231.

[2] Tweya H, Feldacker C, Phiri S, Ben-Smith A, Fenner L, Jahn A, et al. Comparison of treatment outcomes of new smear-positive pulmonary tuberculosis patients by HIV and antiretroviral status in a TB/HIV clinic, Malawi. PLOS ONE 2013;8:e56248. https://doi.org/10.1371/journal.pone.0056248.

[3] Kendall EA, Sahu S, Pai M, Fox GJ, Varaine F, Cox H, et al. What will it take to eliminate drug-resistant tuberculosis? Int J Tuberc Lung Dis 2019;23:535–46. https://doi.org/10.5588/ijtld.18.0217.

[4] World Health Organization. Global tuberculosis report 2021. In: World health organization. Geneva: World Health Organization; 2021. p. 57.

[5] Goletti D, Lindestam Arlehamn CS, Scriba TJ, Anthony R, Cirillo DM, Alonzi T, et al. Can we predict tuberculosis cure? What tools are available? Eur Respir J 2018;52:1801089. https://doi.org/10.1183/13993003.01089-2018.

[6] Zimmer AJ, Lainati F, Aguilera Vasquez N, Chedid C, McGrath S, Benedetti A, et al. Biomarkers that correlate with active pulmonary tuberculosis treatment response: a systematic review and meta-analysis. J Clin Microbiol 2022;60:e0185921. https://doi.org/10.1128/jcm.01859-21.

[7] World Health Organization. Target product profiles: tests for tuberculosis treatment monitoring and optimization. Geneva: World Health Organization; 2023.

[8] FDA-NIH Biomarker Working Group. Monitoring biomarker. In: BEST (Biomarkers EndpointS and other Tools) resource, editor. Food and drug administration. Bethesda: National Institutes of Health (US); 2016.

[9] Heyckendorf J, Georghiou SB, Frahm N, Heinrich N, Kontsevay I, Reimann M, et al. Tuberculosis treatment monitoring and outcome measures: new interest and new strategies. Clin Microbiol Rev 2022;35:e0022721. https://doi.org/10.1128/cmr.00227-21.

[10] World Health Organization. Meeting report of the WHO expert consultation on drug-resistant tuberculosis treatment outcome definitions, 17–19 November 2020. Geneva: World Health Organization; 2021. p. 44.

[11] World Health Organization. Definitions and reporting framework for tuberculosis—2013 revision (updated December 2014 and January 2020). Geneva: World Health Organization; 2013. p. 48.

[12] Linh NN, Viney K, Gegia M, Falzon D, Glaziou P, Floyd K, et al. World Health Organization treatment outcome definitions for tuberculosis: 2021 update. Eur Respir J 2021;58:2100804. https://doi.org/10.1183/13993003.00804-2021.

[13] Global Laboratory Initiative—a working group of the Stop TB Partnership, mycobacteriology laboratory manual. 1st ed. Geneva: Stop TB Partnership; 2014.

[14] Horne DJ, Royce SE, Gooze L, Narita M, Hopewell PC, Nahid P, et al. Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis. Lancet Infect Dis 2010;10:387–94. https://doi.org/10.1016/s1473-3099(10)70071-2.

[15] Peter JG, Theron G, Singh N, Singh A, Dheda K. Sputum induction to aid diagnosis of smear-negative or sputum-scarce tuberculosis in adults in HIV-endemic settings. Eur Respir J 2014;43:185–94. https://doi.org/10.1183/09031936.00198012.

[16] Perez-Velez CM, Marais BJ. Tuberculosis in children. N Engl J Med 2012;367:348–61. https://doi.org/10.1056/NEJMra1008049.

[17] Purohit M, Mustafa T. Laboratory diagnosis of extra-pulmonary tuberculosis (EPTB) in resource-constrained setting: state of the art, challenges and the need. J Clin Diagn Res 2015;9:Ee01–6. https://doi.org/10.7860/jcdr/2015/12422.5792.

[18] Nsubuga R, Adrawa N, Okoboi S, Komuhangi A, Izudi J. Complete sputum smear monitoring among adults with pulmonary tuberculosis in central Uganda: evidence from a retrospective cohort study. BMC Infect Dis 2022;22:191. https://doi.org/10.1186/s12879-022-07178-9.

[19] Izudi J, Tamwesigire IK, Bajunirwe F. Does completion of sputum smear monitoring have an effect on treatment success and cure rate among adult tuberculosis patients in rural Eastern Uganda? A propensity score-matched analysis. PLOS ONE 2019;14:e0226919. https://doi.org/10.1371/journal.pone.0226919.

[20] Bartholomay P, Pelissari DM, de Araujo WN, Yadon ZE, Heldal E. Quality of tuberculosis care at different levels of health care in Brazil in 2013. Rev Panam Salud Publica/Pan Am J Public Health 2016;39:3–11. PMID 27754532.

[21] World Health Organization. Meeting report: high-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, a rapid biomarker-based non-sputum-based test for detecting TB. Geneva: World Health Organization; 2014. p. 11–6.

[22] World Health Organization. Target product profile for next-generation drug-susceptibility testing at peripheral centres. Geneva: World Health Organization; 2021.

[23] Nahid P, Saukkonen J, Mac Kenzie WR, Johnson JL, Phillips PP, Andersen J, et al. Tuberculosis biomarker and surrogate endpoint research roadmap. Am J Respir Crit Care Med 2011;184:972–9. https://doi.org/10.1164/rccm.201105-0827WS.

[24] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

[25] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–30. https://doi.org/10.1056/nejm197810262991705.

[26] Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335–41. https://doi.org/10.1373/clinchem.2005.048595.

[27] Vega V, Rodríguez S, Van der Stuyft P, Seas C, Otero L. Recurrent TB: a systematic review and meta-analysis of the incidence rates and the proportions of relapses and reinfections. Thorax 2021;76:494–502. https://doi.org/10.1136/thoraxjnl-2020-215449.

[28] Guerra-Assunção JA, Houben RM, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis 2015;211:1154–63. https://doi.org/10.1093/infdis/jiu574.

[29] Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med 2013;137:558–65. https://doi.org/10.5858/arpa.2012-0198-RA.

[30] Schumacher SG, Wells WA, Nicol MP, Steingart KR, Theron G, Dorman SE, et al. Guidance for studies evaluating the accuracy of sputum-based tests to diagnose tuberculosis. J Infect Dis 2019;220:S99–107. https://doi.org/10.1093/infdis/jiz258.

[31] Goletti D, Lee MR, Wang JY, Walter N, Ottenhoff THM. Update on tuberculosis biomarkers: from correlates of risk, to correlates of active disease and of cure from disease. Respirology 2018;23:455–66. https://doi.org/10.1111/resp.13272.

[32] Odia T, Malherbe ST, Meier S, Maasdorp E, Kleynhans L, du Plessis N, et al. The peripheral blood transcriptome is correlated with PET measures of lung inflammation during successful tuberculosis treatment. Front Immunol 2020;11:596173. https://doi.org/10.3389/fimmu.2020.596173.

[33] Malherbe ST, Shenai S, Ronacher K, Loxton AG, Dolganov G, Kriel M, et al. Persisting positron emission tomography lesion activity and *Mycobacterium tuberculosis* mRNA after tuberculosis cure. Nat Med 2016;22:1094–100. https://doi.org/10.1038/nm.4177.

[34] Esmail H, Lai RP, Lesosky M, Wilkinson KA, Graham CM, Coussens AK, et al. Characterization of progressive HIV-associated tuberculosis using 2-deoxy-2-[^{18}F]fluoro-D-glucose positron emission and computed tomography. Nat Med 2016;22:1090–3. https://doi.org/10.1038/nm.4161.

[35] Miotto P, Bigoni S, Migliori GB, Matteelli A, Cirillo DM. Early tuberculosis treatment monitoring by Xpert(R) MTB/RIF. Eur Respir J 2012;39:1269–71. https://doi.org/10.1183/09031936.00124711.

[36] Fradejas I, Ontañón B, Muñoz-Gallego I, Ramírez-Vela MJ, López-Roa P. The value of xpert MTB/RIF-generated CT values for predicting the smear status of patients with pulmonary tuberculosis. J Clin Tuberc Other Mycobact Dis 2018;13:9–12. https://doi.org/10.1016/j.jctube.2018.04.002.

[37] Namugenyi J, Musaazi J, Katamba A, Kalyango J, Sendaula E, Kambugu A, et al. Baseline Xpert MTB/RIF ct values predict sputum conversion during the intensive phase of anti-TB treatment in HIV infected patients in Kampala, Uganda: a retrospective study. BMC Infect Dis 2021;21:513. https://doi.org/10.1186/s12879-021-06220-6.

[38] Shenai S, Ronacher K, Malherbe S, Stanley K, Kriel M, Winter J, et al. Bacterial loads measured by the Xpert MTB/RIF assay as markers of culture conversion and bacteriological cure in pulmonary TB. PLOS ONE 2016;11:e0160062. https://doi.org/10.1371/journal.pone.0160062.

[39] Sabiiti W, Azam K, Farmer ECW, Kuchaka D, Mtafya B, Bowness R, et al. Tuberculosis bacillary load, an early marker of disease severity: the utility of tuberculosis Molecular Bacterial Load Assay. Thorax 2020;75:606–8. https://doi.org/10.1136/thoraxjnl-2019-214238.

[40] Ntinginya NE, Bakuli A, Mapamba D, Sabiiti W, Kibiki G, Minja LT, et al. Tuberculosis molecular bacterial load assay reveals early delayed bacterial killing in patients with relapse. Clin Infect Dis 2023;76:e990–4. https://doi.org/10.1093/cid/ciac445.

[41] Walter ND, Born SEM, Robertson GT, Reichlen M, Dide-Agossou C, Ektnitphong VA, et al. *Mycobacterium tuberculosis* precursor rRNA as a measure of treatment-shortening activity of drugs and regimens. Nat Commun 2021;12:2899. https://doi.org/10.1038/s41467-021-22833-6.

[42] Dide-Agossou C, Bauman AA, Ramey ME, Rossmassler K, Al Mubarak R, Pauly S, et al. Combination of *Mycobacterium tuberculosis* RS ratio and CFU improves the ability of murine efficacy experiments to distinguish between drug treatments. Antimicrob Agents Chemother 2022;66:e0231021. https://doi.org/10.1128/aac.02310-21.

[43] Almeida ML, Barbieri MA, Gurgel RQ, Abdurrahman ST, Baba UA, Hart CA, et al. alpha1-acid glycoprotein and alpha1-antitrypsin as early markers of treatment response in patients receiving the intensive phase of tuberculosis therapy. Trans R Soc Trop Med Hyg 2009;103:575–80. https://doi.org/10.1016/j.trstmh.2008.11.024.

[44] Osawa T, Watanabe M, Morimoto K, Okumura M, Yoshiyama T, Ogata H, et al. Serum procalcitonin levels predict mortality risk in patients with pulmonary tuberculosis: a single-center prospective observational study. J Infect Dis 2020;222:1651–4. https://doi.org/10.1093/infdis/jiaa275.

[45] World Health Organization. WHO consolidated guidelines on tuberculosis: module 4: treatment: Drug-resistant tuberculosis treatment. In: World health organization. Geneva: World Health Organization; 2020.

[46] Global Alliance for TB Drug Development. Safety and efficacy of various doses and treatment durations of linezolid plus bedaquiline and pretomanid in participants with pulmonary, XDR-TB, pre-XDR-TB or non-responsive/intolerant MDR-TB (ZeNix) (Clinicaltrials.gov Identifier: NCT03086486), Clinicaltrials.gov. Bethesda: U.S. National Library of Medicine (NLM); 2022.

[47] Medecins Sans Frontieres, Netherlands. Pragmatic clinical trial for a more effective concise and less toxic MDR-TB treatment regimen(s) (TB-PRACTECAL) (Clinicaltrials.gov Identifier: NCT02589782), Clinicaltrials.gov. Bethesda: U.S. National Library of Medicine (NLM); 2022.

[48] Du Bruyn E, Fukutani KF, Rockwood N, Schutz C, Meintjes G, Arriaga MB, et al. Inflammatory profile of patients with tuberculosis with or without HIV-1 co-infection: a prospective cohort study and immunological network analysis. Lancet Microbe 2021;2:e375–85. https://doi.org/10.1016/s2666-5247(21)00037-9.

[49] Togun TO, MacLean E, Kampmann B, Pai M. Biomarkers for diagnosis of childhood tuberculosis: a systematic review. PLOS ONE 2018;13:e0204029. https://doi.org/10.1371/journal.pone.0204029.

[50] Sivro A, McKinnon LR, Yende-Zuma N, Gengiah S, Samsunder N, Abdool Karim SS, et al. Plasma cytokine predictors of tuberculosis recurrence in antiretroviral-treated human immunodeficiency virus-infected individuals from Durban, South Africa. Clin Infect Dis 2017;65:819–26. https://doi.org/10.1093/cid/cix357.

[51] Nahid P, Jarlsberg LG, Kato-Maeda M, Segal MR, Osmond DH, Gagneux S, et al. Interplay of strain and race/ethnicity in the innate immune response to *M. tuberculosis*. PLOS ONE 2018;13:e0195392. https://doi.org/10.1371/journal.pone.0195392.

[52] World Health Organization. WHO consolidated guidelines on tuberculosis, module 4: treatment: Drug-susceptible tuberculosis treatment. In: World health organization. Geneva: World Health Organization; 2022.

[53] Nahid P, Mase SR, Migliori GB, Sotgiu G, Bothamley GH, Brozek JL, et al. Treatment of drug-resistant tuberculosis. An official ATS/CDC/ERS/IDSA clinical practice guideline. Am J Respir Crit Care Med 2019;200:e93–142. https://doi.org/10.1164/rccm.201909-1874ST.

[54] World Health Organization. Lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis of active tuberculosis in people living with HIV. Policy update. In: World health organization. Geneva: World Health Organization; 2019.

[55] Li G, Taljaard M, Van den Heuvel ER, Levine MA, Cook DJ, Wells GA, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol 2017;46:746–55. https://doi.org/10.1093/ije/dyw320.

[56] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527. https://doi.org/10.1136/bmj.h5527.

[57] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6. https://doi.org/10.1136/bmj.39489.470347.AD.