WHO consolidated guidelines on tuberculosis. Module 3: diagnosis – rapid diagnostics for tuberculosis detection

Web Annex D. Evidence synthesis and analysis

Third edition

World Health Organization

WHO consolidated guidelines on tuberculosis. Module 3: diagnosis – rapid diagnostics for tuberculosis detection, third edition. Web Annex D. Evidence synthesis and analysis

ISBN 978-92-4-009047-7 (electronic version)

Suggested citation. Web Annex D. Evidence synthesis and analysis. In: WHO consolidated guidelines on tuberculosis. Module 3: diagnosis – rapid diagnostics for tuberculosis detection, third edition. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO.

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

Sales, rights and licensing. To purchase WHO publications, see https://www.who.int/publications/book-orders. To submit requests for commercial use and queries on rights and licensing, see https://www.who.int/copyright.

This publication forms part of the WHO guideline entitled *WHO consolidated guidelines on tuberculosis. Module 3: diagnosis – rapid diagnostics for tuberculosis detection, third edition*. It is being made publicly available for transparency purposes and information, in accordance with the *WHO handbook for guideline development*, 2nd edition (2014).

The named authors alone are responsible for the views expressed in this publication.

Contents

Acknowledgements

The World Health Organization (WHO) wishes to acknowledge the invaluable contributions made by all the experts who carried out the systematic reviews included in this web annex. Their names and affiliations are included in the sections concerned.

Web Annex D.1. Impact of diagnostic test Xpert MTB/RIF on patient important outcomes for tuberculosis: a systematic review

Frederick Haraka, Eleanor Ochodo, Klaus Reither, and Karen R Steingart

Sections of this report are excerpted from a pre-peer review version of a Cochrane Review: **Impact of diagnostic test Xpert MTB/RIF® on health outcomes for tuberculosis**

Frederick Haraka[1,2,3], Mwaka Kakolwa[4], Samuel G Schumacher[5], Ruvandhi R Nathavitharana[6], Claudia M Denkinger[5,7], Sebastien Gagneux[2,3], Klaus Reither[1,2,3,] Amanda Ross [2,3]

[1]Ifakara Health Institute, Bagamoyo, United Republic of Tanzania
[2]Swiss Tropical and Public Health Institute, Basel, Switzerland
[3]University of Basel, Basel, Switzerland
[4]Ifakara Health Institute, Dodoma, United Republic of Tanzania
[5]Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland
[6]Division of Infectious Diseases, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, United States of America (USA)
[7]Division of Tropical Medicine, Center of Infectious Diseases, University Hospital of Heidelberg, Heidelberg, Germany

Corresponding author:
Frederick Haraka
Ifakara Health Institute
Bagamoyo, United Republic of Tanzania
E-mail: fharaka@ihi.or.tz

BACKGROUND

In 2018, tuberculosis (TB) was associated with 1.2 million deaths and a further 251,000 deaths from tuberculosis disease among people living with HIV (WHO Global tuberculosis report 2019). The absolute number of TB deaths among HIV-negative people fell by 27% between 2000 and 2018, from an estimated 1.7 million in 2000 to 1.2 million in 2018, and similarly the mortality rate fell by 42% (including 3.6% between 2017 and 2018). Among HIV-positive people, the number of TB deaths fell faster, from 624 000 in 2000 to 251,000 in 2018 (a reduction of 60%), and the mortality rate fell by 68% (from 10 to 3.3 per 100,000 population) (WHO Global tuberculosis report 2019). Of the WHO regions, Africa had the highest mortality rate (18%) (WHO Global tuberculosis report 2019). There has been progress in treatment success (cure and treatment completion). Latest data show a global success rate of 85% among new TB cases in 2017 compared to 81% in the previous year (WHO Global tuberculosis report 2019). Overall loss to follow up were high in the WHO region of the Americas accounting for 25%.

Xpert MTB/RIF was endorsed by the WHO in 2010 and since then it has been included in more than 10 high burden countries in their national policies (Cazabon 2016). By 2016 approximately 23 million Xpert cartridge were procured in the public sector in 130 countries under concessional price and approximately 34.4 million overall globally (Cazabon 2016; Cazabon 2017).

The aim of this Cochrane Review was to assess the impact of diagnostic strategies using Xpert MTB/RIF compared to strategies using smear microscopy on people important outcomes. In this WHO report, we considered the following outcomes: all-cause mortality, pre-treatment loss to follow-up, cure, time to diagnosis, and time to treatment initiation.

METHODS

Search methods

We searched the following databases, without language restriction, from 2007 to 27 February 2018 and updated our search from 2017 to 31 July 2019: Cochrane Infectious Disease Group (CIDG) Specialized Register; Cochrane CentralRegister of Controlled Trials (CENTRAL), published in the Cochrane Library;MEDLINE OVID; Embase OVID; CINAHL EBSCO; LILACS (Latin American andCaribbean Health Science Information database; BIREME); Science Citation IndexExpanded (Web of Science), Social Sciences citation index (Web of Science), andConference Proceedings Citation Index - Social Science & Humanities (Web ofScience). We also searched the WHO International Clinical Trials Registry Platform(www.who.int/ictrp/search/en/), ClinicalTrials.gov (clinicaltrials.gov/), and the Pan AfricanClinical Trials Registry (www.pactr.org/) to identify ongoing trials using (tuberculosis OR TB) AND (Xpert orGeneXpert or "sputum microbiology" or "sputum microscopy") as search terms.

Selection criteria

We included published randomized controlled trials and cluster randomized trials that compared the use of Xpert MTB/RIF and smear microscopy on health outcomes. We only included trials if they evaluated expectorated sputum consistent with routine practice. Multiple publications of the trial were included only once by including the publication with the largest sample size for the outcomes assessed and the most

detailed information. Except for analysis of the outcome on cure, we included another publication by Trajman A 2015 from the trial in Brazil by Durovni 2014. We excluded studies on accuracy of Xpert MTB/RIF and those without a comparison group.

Data collection and analysis

Two review authors independently extracted data using a piloted data extraction tool. We resolved disagreements through discussion or by consulting a third review author. We extracted the following data: study details (first author, year of publication), participant details, intervention, control, outcome measured and how it was measured, covariates, length of follow-up, and measure of effect with 95% confidence intervals. For binary outcomes, we extracted the relative risk or odds ratio if available. For time-to-event outcomes, we extracted the log hazard ratio (HR) with standard error or confidence interval.

In cluster randomized trials, we recorded the number of participants and clusters randomised to each diagnostic arm, and the number of participants monitored for each outcome of interest and the number of events in each diagnostic arm. For cluster-randomized trials that were adjusted for clustering, we extracted the adjusted measures of effect for each outcome and method of adjustment. In studies that are not adjusted for clustering, we extracted the number of clusters randomised or the mean cluster size and the intraclass correlation coefficient (ICC), if available. We also extracted data relevant for the assessment of the risk of bias.

RESULTS

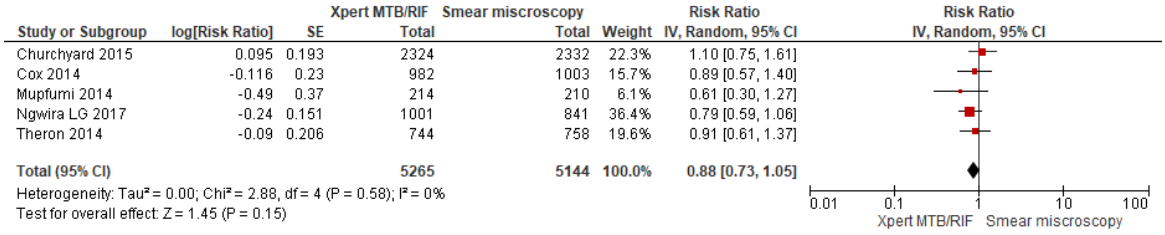
For the impact of Xpert MTB/RIF on all-cause mortality we included five studies (10,409 participants). We included two individually randomized trials (Mupfumi 2014; Theron 2014) and three cluster randomised trials (Churchyard 2015; Cox 2014; Ngwira LG 2017). All studies were conducted in high TB burden and high TB/HIV burden countries. There were two trials in South Africa (Churchyard 2015; Cox 2014), one in Zimbabwe (Mupfumi 2014), one in Malawi (Ngwira LG 2017) and one multi-country study with sites in South Africa, United Republic of Tanzania and Zimbabwe (Theron 2014). All studies were conducted in outpatient settings and enrolled participants aged 18 years or older.

PICO 1 Subquestion 15: Should Xpert MTB/RIF vs smear microscopy be used in adults with signs and symptoms of pulmonary tuberculosis?

Impact on all-cause mortality

Overall, all-cause mortality occurred in 248 (4.7%) of 5265 in the Xpert MTB/RIF group and 292 (5.7%) of 5144 in smear group. Compared to smear microscopy, the overall risk of mortality was estimated to be RR 0.88 (0.73, 1.05) (5 trials, 10,409 participants; **moderate-certainty evidence**), (Cox 2014; Churchyard 2015; Ngwira LG 2017 Mupfumi 2014; Theron 2014).

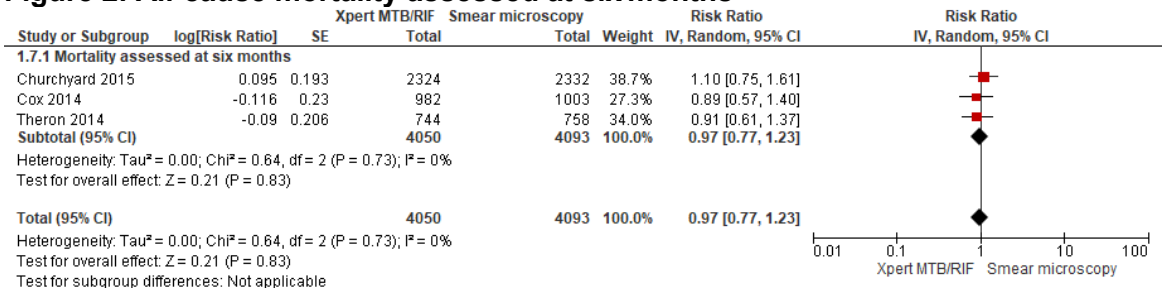
Figure 1: Overall impact on all-cause mortality



Mortality assessed at six months

Restricting the analysis to the three studies that assessed mortality at six months only the risk of all-cause mortality was estimated to be RR 0.97 (95%CI: 0.77, 1.23), (3 trials, 8143 participants), (Churchyard 2015; Cox 2014; Theron 2014).

Figure 2: All-cause mortality assessed at six months



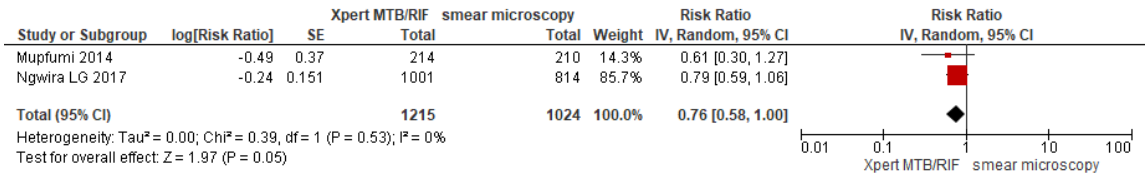
Beyond treatment period

Only one study (Ngwira LG 2017) assessed mortality beyond the treatment period. All-cause mortality was assessed at 12 months among HIV-positive patients. Overall, all-cause mortality was 22% lower in the Xpert MTB/RIF group, 55 deaths of 818 person-years (6.7 per 100 person-years), compared to the smear microscopy group, 58 deaths of 685 person-years (8.6 per 100 person-years), RR 0.78 (0.58, 1.06). In a sub-group analysis among patients with advanced AIDS, all-cause mortality was lower in the Xpert MTB/RIF group, 32 deaths of 231 person-years (13.9 per 100 person-years), compared to the smear microscopy group, 36 deaths of 127 person-years (28.3 per 100 person-years) RR 0.43 (0.22, 0.87).

In HIV-positive participants

In studies that included only HIV-positive participants (Ngwira LG 2017; Mupfumi 2014), the risk of all-cause mortality was estimated to be RR 0.76 (95% CI 0.58, 1.00), (2 trials, 2239 participants), **(moderate-certainty evidence**).

Figure 3: Impact on all-cause mortality in people living with HIV



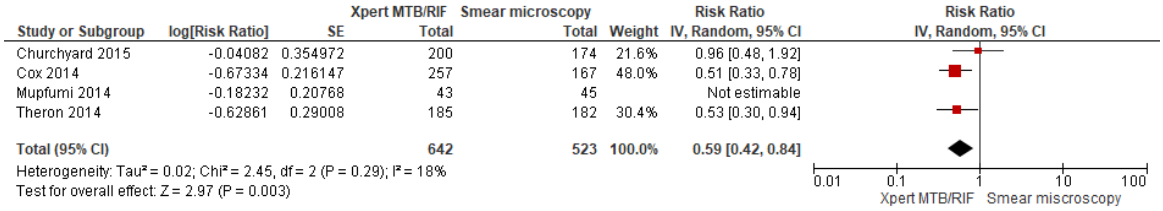
Summary of individual patient data meta-analysis

Di Tanna 2019 compared the effect of Xpert MTB/RIF and smear microscopy on all-cause mortality using patient level data. In the analysis for mortality, three randomized trials were included (Churchyard 2015; Cox 2014: Theron 2014). All three trials were conducted in South Africa and assessed mortality at six months among adults in outpatient settings. In summary, for the primary outcome of 6-month mortality risk among outpatients tested for TB, overall all cause 6-month mortality occurred in 182 (4.5%) of 4050 patients in the Xpert MTB/RIF group and 217 (5.3%) of 4093 patients in the sputum smear microscopy group, pooled adjusted OR 0.88 (95% CI 0.68, 1.14); P = 0.34, (3 trials, 8143 participants). A stratified analysis among HIV-positive individuals showed a pooled adjusted OR of 0.83 (0.65, 1.05); P = 0.12.

Pre-treatment loss to follow up

Xpert MTB/RIF was found to reduce the risk of pre-treatment loss to follow-up with an estimated RR of 0.59 (0.42, 0.84) (3 trials,1165 participants; moderate-certainity evidence), (Churchyard 2015; Cox 2014; Theron 2014).

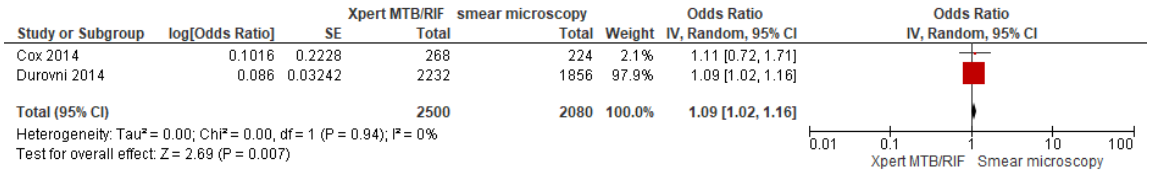
Figure 4: Proportion of pre-treatment loss to follow up



Cure

Included studies assessed treatment success as an outcome after six months of TB treatment. The WHO defines treatment success for TB if there is evidence of cure and that cure is confirmed bacteriologically or if the patient has completed treatment without evidence of treatment failure. For analysis of this outcome, we considered treatment completion as a proxy for cure hence we defined 'cure' in those cured and those completing treatment. In comparison to smear microscopy, Xpert MTB/RIF significantly increased the odds of cure, OR 1.09 (1.02, 1.16), (2 trials, 4580 participants; **high-certainty evidence**), (Cox 2014; Durovni 2014).

Figure 5: Proportion cured



Time to diagnosis

The meta-analysis included two trials involving 1924 participants (Theron 2014; Mupfumi 2014). The median time to diagnosis was 0.5 days ((Interquartile range (IQR) 0.5 to 10 days)) for each group, pooled HR 1.05 (95% CI 0.93, 1.19); P = 0.43, (Di Tanna 2019).

The results were in the direction of benefit and we judged this analysis as **high-certainty evidence**.

Time to treatment
The meta-analysis included four trials involving 8208 participants (Theron 2014; Mupfumi 2014; Cox 2014; Churchyard 2015). The median time to treatment was four days (IQR 1 to 10 days) for the Xpert MTB/RIF group and five days (IQR 1 to 15 days) for the smear microscopy group, pooled HR 1.0 (95% CI 0.75, 1.32); P = 0.99, (Di Tanna 2019). We judged this analysis as **moderate-certainty evidence.**

GRADE tables and certainty of evidence
For the main outcome (mortality), we judged risk of bias to be low across studies. There was no evidence of inconsistency or indirectness. We considered imprecision to be serious and downgraded one level. Explanations for GRADE-ing are shown below in the GRADE tables.

AUTHORS' CONCLUSIONS

We found that, compared with smear microscopy, the use of Xpert MTB/RIF did not result in a statistically significant reduction in all cause mortality. However, we caution about interpreting non-significance to mean no effect when the 95% confidence interval likely includes an effect that may be clinically important. We found that the use of Xpert MTB/RIF decreased pre-treatment loss to follow-up and increased the proportion of TB patients cured.

Table 1: Descriptive summary of studies included for all-cause mortality assessment.

Study, year	Country	Design	Settings	Sample size	Month mortality assessed	No. of patients tested in smear group	No. (%) deaths in smear group	No. of patients tested in Xpert group	No. (%) deaths in Xpert group	RR	P value
Churchyard 2015	South Africa	Cluster RCT	Outpatient -primary healthcare clinics	4656	6	2332	116 (5)	2324	91 (3.9)	1.1 (0.75-1.62)	0.61
Cox 2014	South Africa	Cluster RCT	Outpatient -primary healthcare clinics	1985	6	1003	38 (3.8)	982	33 (3.4)	0.89 (0.58-1.75)	0.51
Mupfumi 2014	Zimbabwe	RCT	Outpatient - specialize d infectious disease clinic	424	3	214	17(9.9)	210	11 (6)	0.61 (0.29-1.27)	0.19
Ngwira 2017	Malawi	Cluster RCT	Outpatient -HIV primary healthcare clinics	1842	12	841	58 (8.6 per 100 person-years)	1001	55 (6.7 per 100 person- years)	0.78 (0.58-1.06)	0.1
Theron 2014	South Africa, United Republic of Tanzania, Zambia,	RCT	Outpatient -primary healthcare clinics	1502	6	758	63 (8)	744	58 (8)	*AoR 0.92 (0.61-1.39)	0.7

Odds ratio was included given that odds ratio approximates risk ratio if outcomes are rare e.g. Mortality and inclusion of odds would not lead to overestimation of the measure of effect.

Table 2: Descriptive summary of studies included for pre-treatment loss to follow up

Study, year	Country	Design	Settings	Total number tested positive for TB	Month pre-treatment loss to follow up assessed	Number of patients tested for TB in smear group	Number lost to follow up before treatment initiation in smear group N (%)	Number of patients tested for TB in Xpert group	Number lost to follow up before treatment initiation in Xpert group N (%)	RR	P value
Churchyard 2015	South Africa	Cluster RCT	Outpatient- primary healthcare clinics	374	1	174	26 (15)	200	34 (17)	0.96 (0.48-1.93)	0.91
Cox 2014	South Africa	Cluster RCT	Outpatient- primary healthcare clinics	424	3	167	41 (25)	257	32(13)	0.51 (0.33-0.77)	0.0052
Theron 2014	South Africa, United Republic of Tanzania, Zambia, Zimbabwe	RCT	Outpatient primary healthcare clinics	367	6	182	28 (15)	185	15 (8)	-	0.03

Table 3: Descriptive summary of studies included for cure

Study, year	Country	Design	Settings	Total number treated for TB	Number of patients treated for TB in smear group	Number cured in smear group N (%)	Number of patients treated for TB in Xpert group	Number lost to follow up before treatment initiation in Xpert N (%)	OR	P value
Cox 2014	South Africa	Cluster RCT	Outpatient-primary healthcare clinics	492	224	176 (78.6)	268	215 (80.2)	-	0.75
Durovni 2014	Brazil	Step wedged cluster	Outpatient primary healthcare clinics	4088	1856	1267 (68.3)	2232	1571 (70.4)	-	-

Table 4: Descriptive summary of studies included for time to diagnosis

Study , year	Country	Design	Settings	Sample size	Number of patients in smear arm	Number of patients in Xpert arm	Time to diagnosis smear arm, median days, (IQR)	Time diagnosis Xpert arm, median days, (IQR)	HR (95% CI)	P value
Theron 2014	South Africa, United Republic of Tanzania, Zambia, and Zimbabwe	RCT	Outpatient-primary healthcare clinics	1500	757	743	0.5 (na)	0.5 (na)	-	-
Mupfumi 2014	Zimbabwe	RCT	Outpatient-specialized infectious disease clinic	424	210	214	6 (1-25)	2 (1-13)	-	-
Di Tanna 2019	South Africa, United Republic of Tanzania, Zambia,	IPD meta-analyses	Outpatient health care clinics	1924	967	957	0.5 (0.5-10)	0.5 (0.5-10)	1.05 (0.93-1.19)	0.43

Table 5: Descriptive summary of studies included for time to treatment

Study ID	Country	Design	Settings	Sample size	Number of patients in smear arm	Number of patients in Xpert arm	Time to diagnosis smear arm, median days, (IQR)	Time diagnosis Xpert arm, median days, (IQR)	HR (95% CI)	P value
Theron 2014	South Africa, United Republic of Tanzania, Zambia, and Zimbabwe	RCT	Outpatient-primary healthcare clinics	1219	643	576	1 (0-4)	0 (0-3)	-	-
Mupfumi 2014	Zimbabwe	RCT	Outpatient-specialized infectious disease clinic	422	210	212	8 (3-23)	5 (3-13)	-	-
Cox 2014	South Africa	RCT	Outpatient-primary healthcare clinic	1911	968	943	8 (2-22)	4 (2-10)	-	-
Churchyard 2015	South Africa	RCT	Outpatient-primary healthcare clinics	4656	2332	2324	10 (na)	7 (na)	-	-
Di Tanna 2019	South Africa, United Republic of Tanzania, Zimbabwe	IPD meta-analyses	Outpatient health care clinics	8208	4153	4055	5 (1-15)	4 (1-10)	1.00 (0.75-1.32)	0.99

Web Annex D.2. Xpert MTB/RIF and Xpert Ultra for detecting active tuberculosis in adults with signs and symptoms of pulmonary TB: an updated systematic review

David J Horne[1], Jerry S Zifodya[2,] Mikashmi Kohli[3], Ian Schiller[4], Nandini Dendukuri[4], Deanna Tollefson[5], Samuel G Schumacher[6], Eleanor A Ochodo[7], Madhukar Pai[3], Karen R Steingart[8]

[1]Department of Medicine, Division of Pulmonary, Critical Care, and Sleep Medicine, University of Washington, Seattle, WA, United States of America
[2]Department of Medicine, Section of Pulmonary Diseases, Critical Care and Environmental Medicine, Tulane University School of Medicine, New Orleans, LA, USA
[3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[4]Division of Clinical Epidemiology, McGill University Health Centre - Research Institute, Montreal, Canada
[5]Department of Global Health, University of Washington, Seattle, WA, USA
[6]FIND, Geneva, Switzerland
[7]Centre for Evidence-based Health Care, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[8]Honorary Research Fellow, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom of Great Britain and Northern Ireland, United Kingdom

Corresponding author:
David J Horne, MD, MPH
Harborview Medical Center
325 Ninth Ave, 359762
Seattle, WA 98104 USA
dhorne@uw.edu

BACKGROUND

Xpert MTB/RIF and Xpert Ultra are WHO-recommended rapid tests that simultaneously detect tuberculosis and rifampicin resistance in people with signs and symptoms of tuberculosis and are suitable for use at lower levels of the health system. This systematic review assessed the diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for detecting tuberculosis and rifampicin resistance from pulmonary specimens in adults. There were an estimated 10 million incident cases of tuberculosis in 2018 and of the 7 million reported cases, 85% involved the lungs (WHO Global Tuberculosis Report 2019). In 2018, there were about half a million new cases of rifampicin-resistant TB, and of these, 78% had multidrug-resistant TB (WHO Global Tuberculosis Report 2019). A previous Cochrane Review found Xpert MTB/RIF sensitive and specific for pulmonary tuberculosis, although sensitivity was decreased in paucibacillary samples (Steingart 2014). We performed a systematic review update for a WHO policy update, as additional Xpert MTB/RIF studies have been published and Xpert Ultra introduced since the prior systematic review.

METHODS

Search methods

We searched the Cochrane Infectious Diseases Group Specialized Register, MEDLINE, Embase, Science Citation Index, Web of Science, Latin American Caribbean Health Sciences Literature, Scopus, the WHO International Clinical Trials Registry Platform, the International Standard Randomized Controlled Trial Number Registry, and ProQuest, to 11 October 2018 for studies evaluating Xpert MTB/RIF and to 23 September 2019 for studies evaluating Xpert Ultra, without language restriction.

Selection criteria

We included randomized trials, cross-sectional, and cohort studies using respiratory specimens that evaluated Xpert MTB/RIF, Xpert Ultra, or both against the reference standards of culture for tuberculosis and culture-based drug susceptibility testing or MTBDR*plus* for rifampicin resistance. For Xpert Ultra, we also included a composite reference standard that included clinical components as defined by the primary study authors. Only studies that enrolled adults (≥15 years of age) were eligible. For the evaluation of TB detection, we included studies that evaluated the index tests in people with signs and symptoms of pulmonary tuberculosis except for studies in people living with HIV, for whom studies were eligible *irrespective of* signs and symptoms of pulmonary tuberculosis (e.g. studies that performed tuberculosis screening in people living with HIV as part of intensified case finding or prior to TB preventive therapy).

For tuberculosis detection, we included studies that reported data comparing the index test(s) to an acceptable reference standard from which we could extract true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. For the detection of rifampicin resistance, we included studies that allowed estimation of sensitivity and specificity (i.e. if no rifampicin resistance was detected by the reference test in any of the specimens, then the study was excluded).

Data collection and analysis

Four review authors independently assessed studies for eligibility. Working in pairs, the review authors extracted data using a standardized form. When possible, we extracted data by sputum smear and HIV status. In addition, for Xpert Ultra studies, we extracted data by history of prior tuberculosis and results of repeat testing when initial specimens were "trace" positive. We assessed methodological quality using QUADAS-2 and performed meta-analyses to estimate pooled sensitivity and specificity separately for Xpert MTB/RIF and Xpert Ultra and separately for pulmonary tuberculosis and rifampicin resistance. We investigated potential sources of

heterogeneity by reference standard and clinical subgroup. Most analyses used a bivariate random-effects model. For pulmonary tuberculosis detection, we first estimated accuracy using all included studies and then only the subset of studies where participants were unselected, i.e. not selected based on prior microscopy testing or prior history of tuberculosis.

RESULTS

Detection of pulmonary tuberculosis

For detection of pulmonary tuberculosis, we identified a total of 94 studies. The total includes one study that provided data for two cohorts and we classified these as two distinct studies, Mishra 2019a and Mishra 2019b. A total of 85 studies (40,652 participants) evaluated Xpert MTB/RIF and nine studies (3881 participants) evaluated both Xpert Ultra and Xpert MTB/RIF.

Of the total 94 studies, 50 (53%) took place in high tuberculosis burden and 54 (57%) in high TB/HIV burden countries. Most studies had low risk of bias. Regarding applicability, most studies had low concern about applicability because participants in these studies were evaluated in primary care facilities, local hospitals, or both settings.

PICO 1, pulmonary TB in adults

1. Should Xpert MTB/RIF be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity (95% Credible Interval (CrI)) were 85% (82 to 88%) and 98% (97 to 98%), respectively (70 studies, 37,237 unselected participants; high-certainty evidence).

For a population of 1000 people where 100 have pulmonary tuberculosis on culture, 103 would be Xpert MTB/RIF-positive and 18 (17%) would not have tuberculosis (false-positives); 897 would be Xpert MTB/RIF-negative and 15 (2%) would have pulmonary tuberculosis (false-negatives).

2. Should Xpert Ultra be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert Ultra pooled sensitivity and specificity (95% CrI) were 90% (84% to 94%) and 96% (93% to 98%), respectively (6 studies, 2654 unselected participants; high-certainty evidence).

For a population of 1000 people where 100 have pulmonary tuberculosis on culture, 130 would be Xpert Ultra-positive and 40 (31%) would not have pulmonary tuberculosis (false-positives); 870 would be Xpert Ultra-negative and 10 (1%) would have pulmonary tuberculosis (false-negatives).

3. Should Xpert Ultra be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB, against a composite reference standard?

Xpert Ultra sensitivity and specificity ranges were 80% to 96% and 96% to 100%, respectively (2 studies, 433 unselected participants; very low-certainty evidence for sensitivity; low-certainty evidence for specificity). We did not perform a meta-analysis owing to insufficient data.

4. Should Xpert MTB/RIF vs. Xpert Ultra be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity (95% CrI) were 83% (76% to 88%) and 99% (97% to 99%), respectively (6 studies, 2654 unselected participants; high-certainty evidence).
Xpert Ultra pooled sensitivity and specificity (95% CrI) were 90% (84% to 94%) and 96% (93% to 98%), (6 studies, 2654 unselected participants; high-certainty evidence).

Xpert MTB/RIF		Xpert Ultra	
Sensitivity	0.83 (95% CrI: 0.76 to 0.88)	Sensitivity	0.90 (95% CrI: 0.84 to 0.94)
Specificity	0.99 (95% CrI: 0.97 to 0.99)	Specificity	0.96 (95% CI: 0.93 to 0.97)

For a population of 1000 people where 100 have pulmonary tuberculosis on culture, 96 would be Xpert MTB/RIF-positive and 13 (14%) would not have pulmonary tuberculosis (false-positives) while 130 would be Xpert Ultra-positive and 40 (31%) would not have pulmonary tuberculosis (false-positives). 904 would be Xpert MTB/RIF-negative and 17 (2%) would have pulmonary tuberculosis (false-negatives) while 870 would be Xpert Ultra-negative and 10 (1%) would have pulmonary tuberculosis (false-negatives).

5. Should Xpert MTB/RIF be used to diagnose pulmonary TB in smear-positive adults with signs and symptoms, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity (95% CrI) was 98% (97% to 98%) (45 studies, 4064 participants; high-certainty evidence).[1]

6. Should Xpert MTB/RIF be used to diagnose pulmonary TB in smear-negative adults with signs and symptoms, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity (95% CrI) were 67% (62% to 72%) and 98% (98% to 99%), respectively (45 studies, 18,962 participants; high-certainty evidence).

For a population of 1000 people where 100 have smear-negative, culture positive pulmonary tuberculosis, 85 would be Xpert MTB/RIF-positive and 18 (21%) would not have pulmonary tuberculosis (false-positives); 915 would be Xpert MTB/RIF-negative and 33 (4%) would have pulmonary tuberculosis (false-negatives).

7. Should Xpert Ultra be used to diagnose pulmonary TB in smear-positive adults with signs and symptoms, against a microbiological reference standard?
Xpert Ultra pooled sensitivity (95% CrI) was 99% (98% to 100%) (6 studies, 575 participants; high-certainty evidence).

8. Should Xpert Ultra be used to diagnose pulmonary TB in smear-negative adults with signs and symptoms, against a microbiological reference standard?
Xpert Ultra pooled sensitivity and specificity (95% CrI) were 79% (70% to 87%) and 94% (87% to 97%), respectively (7 studies, 2547 participants; high-certainty evidence).

For a population of 1000 people where 100 have smear-negative, culture positive pulmonary tuberculosis, 136 would be Xpert Ultra-positive and 57 (42%) would not have pulmonary tuberculosis (false-positives); 864 would be Xpert Ultra-negative and 21 (2%) would have pulmonary tuberculosis (false-negatives).

[1] We performed a univariate analysis for sensitivity. We could not estimate Xpert MTB/RIF pooled specificity in smear-positive adults because, in many studies, there were few or zero false positive and true negative values reported.

9. Should Xpert MTB/RIF be used to diagnose pulmonary TB in HIV-positive adults with signs and symptoms, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity (95% CrI) were 81% (75% to 86%) and 98% (97% to 99%), respectively (14 studies, 4664 unselected participants; high-certainty evidence).

For a population of 1000 HIV-positive people where 100 have pulmonary tuberculosis on culture, 99 would be Xpert MTB/RIF-positive and 18 (18%) would not have pulmonary tuberculosis (false-positives); 901would be Xpert MTB/RIF-negative and 19 (2%) would have pulmonary tuberculosis (false-negatives).

10. Should Xpert Ultra be used to diagnose pulmonary TB in HIV-positive adults with signs and symptoms, against a microbiological reference standard?
Xpert Ultra pooled sensitivity and specificity (95% CrI) were 87% (74% to 94%) and 92% (79% to 96%), respectively (3 studies, 627 unselected participants; low-certainty evidence).

For a population of 1000 HIV-positive people where 100 have pulmonary tuberculosis on culture, 160 would be Xpert Ultra-positive and 73 (46%) would not have pulmonary tuberculosis (false-positives); 840 would be Xpert Ultra-negative and 13 (2%) would have pulmonary tuberculosis (false-negatives).

11. Should Xpert MTB/RIF be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB with prior TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity (95% CrI) were 82% (74% to 88%) and 96% (93% to 98%), respectively (11 studies, 4196 unselected participants; moderate-certainty evidence for sensitivity; high-certainty evidence for specificity).

For a population of 1000 people with prior TB where 100 have pulmonary tuberculosis on culture, 118 would be Xpert MTB/RIF-positive and 36 (31%) would not have pulmonary tuberculosis (false-positives); 882 would be Xpert MTB/RIF-negative and 18 (2%) would have pulmonary tuberculosis (false-negatives).[2]

12. Should Xpert Ultra be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB with prior TB, against a microbiological reference standard?
Xpert Ultra pooled sensitivity and specificity (95% CrI) were 84% (72% to 91%) and 86% (72% to 94%), respectively (4 studies, 602 unselected participants; low-certainty evidence).

For a population of 1000 people with prior TB where 100 have pulmonary tuberculosis on culture, 206 would be Xpert Ultra-positive and 122 (59%) would not have pulmonary tuberculosis (false-positives); 794 would be Xpert Ultra-negative and 16 (2%) would have pulmonary tuberculosis (false-negatives).[3]

[2] This analysis included studies in which ≥25% of participants had a history of prior tuberculosis.

[3] This analysis included only participants with a history of prior tuberculosis.

Table. PICO 1, diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for pulmonary TB and rifampicin resistance in adults

PICO sub-question	Test, analysis group	Reference	Pooled Sensitivity % (95% CrI)	Pooled Specificity % (95% CrI)	Positive Predictive Value % (95% CI)	Negative Predictive Value % (95% CI)
1	Xpert MTB/RIF, unselected	MRS	85 (82 to 88)	98 (97 to 98)	83 (75 to 83)	98 (98 to 99)
2	Xpert Ultra, unselected	MRS	90 (84 to 94)	96 (93 to 98)	69 (56 to 81)	99 (98 to 99)
3	Xpert Ultra	CRS	-	-	-	-
4*	Xpert MTB/RIF vs Ultra, direct comparison	MRS	83 (76 to 88)	99 (97 to 99)	86 (76 to 94)	98 (97 to 99)
5	Xpert MTB/RIF, smear positive	MRS	98 (97 to 98)	Did not estimate	Could not estimate	Could not estimate
6	Xpert MTB/RIF, smear negative	MRS	67 (62 to 72)	98 (98 to 99)	79 (78 to 89)	96 (96 to 97)
7	Xpert Ultra, smear positive	MRS	99 (98-100)	Did not estimate	Could not estimate	Could not estimate
8	Xpert Ultra, smear negative	MRS	79 (70 to 87)	94 (87 to 97)	58 (38 to 76)	98 (96 to 99)
9	Xpert MTB/RIF, HIV positive	MRS	81 (75 to 86)	98 (97 to 99)	82 (74 to 91)	98 (97 to 98)
10	Xpert Ultra, HIV positive	MRS	87 (74 to 94)	92 (79 to 96)	54 (29 to 75)	98 (96 to 99)
11	Xpert MTB/RIF, prior TB history	MRS	82 (74 to 88)	96 (93 to 98)	69 (54 to 83)	98 (97 to 99)
12	Xpert Ultra, prior TB history	MRS	84 (72 to 91)	86 (72 to 94)	41 (22 to 63)	98 (96 to 99)
13	Xpert MTB/RIF for rifampicin resistance	Culture-based DST, LPA	96 (94 to 97)	98 (98 to 99)	84 (84 to 92)	99 (99 to 100)
14	Xpert Ultra for rifampicin resistance	Culture-DST, LPA	94 (87 to 98)	99 (98 to 100)	91 (80 to 97)	99 (99 to 100)

Predictive values were determined at pre-test probability of 10%. Dashes indicate analyses where data were insufficient to perform meta-analyses. Abbreviations: CI: Confidence interval; CrI: Credible interval; CRS: composite reference standard; DST: drug susceptibility testing; LPA: line probe assay; MRS: microbiological reference standard (culture); *Results in this row are for Xpert MTB/RIF. Results for Xpert Ultra (direct comparison) are in PICO subquestion 2.

Detection of rifampicin resistance

For detection of rifampicin resistance, a total of 57 studies (8287 participants) evaluated Xpert MTB/RIF and eight studies (1039 participants) evaluated Xpert Ultra. The total number of Xpert Ultra studies includes one study that provided data for two cohorts and we classified these as two distinct studies, Mishra 2019a and Mishra 2019b. Of the 57 studies, 27 took place in high MDR-TB burden countries. We judged most studies as having low risk of bias.

13. Should Xpert MTB/RIF in sputum be used to diagnose rifampicin resistance in adults with signs and symptoms of pulmonary TB?

Xpert MTB/RIF pooled sensitivity and specificity (95% CrI) were 96% (94 to 97) and 98% (98 to 99), respectively (48 studies, 8020 participants; high-certainty evidence).

For a population of 1000 people where 100 have rifampicin-resistant tuberculosis, 114 would be Xpert MTB/RIF-positive and 18 (16%) would not have rifampicin resistance (false-positives); 886 would be Xpert MTB/RIF-negative and 4 (0.4%) would have rifampicin resistance (false-negatives).

14. Should Xpert Ultra in sputum be used to diagnose rifampicin resistance in adults with signs and symptoms of pulmonary TB?
Xpert Ultra pooled sensitivity and specificity were 94% (87% to 98%) and 99% (98% to 100%), respectively (5 studies, 930 unselected participants; high-certainty evidence).

For a population of 1000 people where 100 have rifampicin-resistant tuberculosis, 103 would be Xpert Ultra-positive and 9 (9%) would not have rifampicin resistance (false-positives); 897 would be Xpert Ultra-negative and 6 (1%) would have rifampicin resistance (false-negatives).

PICO 5
61. Should Xpert Ultra repeated test be used to diagnose pulmonary TB in adults with signs and symptoms of pulmonary TB who have an initial Ultra trace result, against a microbiological reference standard?
We identified three studies: Mishra 2019a (4 participants), Piersimoni 2019 (4 participants), and Dorman 2018 (42 participants). Dorman 2018 and Piersimoni 2019 retested the same initial samples. Mishra 2019a retested only those participants with discrepant (Ultra trace positive/culture negative) results and re-testing was performed on new specimens obtained a median of 444 days (range 245 to 526 days) after initial testing. Xpert Ultra accuracy in Mishra 2019a and Piersimoni 2019 was 100% for both sensitivity and specificity. Dorman 2018 found sensitivity of 69% (95% CI 39 to 91) and specificity of 66% (46 to 82). We judged certainty of evidence as very low and did not perform a meta-analysis owing to the limited number of studies.

AUTHORS' CONCLUSIONS

We found Xpert MTB/RIF to be sensitive and specific for diagnosing pulmonary tuberculosis and rifampicin resistance, consistent with findings reported previously. Xpert MTB/RIF was more sensitive for tuberculosis in smear-positive than smear-negative participants and HIV-negative than HIV-positive participants.

- In head-to-head comparisons, Xpert Ultra had higher sensitivity (90%) than Xpert MTB/RIF (83%) and lower specificity (96%) than Xpert MTB/RIF (99%) for tuberculosis detection (6 studies).
- Among people living with HIV, Xpert Ultra again demonstrated higher sensitivity (87%) than MTB/RIF (81%) and lower specificity (92%) than Xpert MTB/RIF (98%) for tuberculosis detection.
- The greatest difference in sensitivity between Xpert Ultra and Xpert MTB/RIF was with smear-negative specimens (79% and 67%, respectively), (indirect comparison).
- Among participants with a history of prior tuberculosis, Xpert Ultra (86%) specificity was lower than that of Xpert MTB/RIF (96%), (indirect comparison).
- Limited data were available to inform decisions on the re-testing of Ultra trace positive specimens.
- Xpert MTB/RIF and Xpert Ultra had similar sensitivity and specificity for rifampicin resistance.

Xpert MTB/RIF and Xpert Ultra provide accurate results and can allow rapid initiation of treatment for multidrug-resistant tuberculosis.

Web Annex D.3 Xpert MTB/RIF and Xpert Ultra for detecting active tuberculosis in adults with signs and symptoms of extra-pulmonary TB: an updated systematic review

Mikashmi Kohli[1], Ian Schiller[2], Nandini Dendukuri[2], Keertan Dheda[3], Claudia M Denkinger[4], Samuel G Schumacher[5], Karen R Steingart[6]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[2]Division of Clinical Epidemiology, McGill University Health Centre - Research Institute, Montreal, Canada
[3]Centre for Lung Infection and Immunity Unit, Department of Medicine and UCT Lung Institute, University of Cape Town, Cape Town, South Africa
[4] Centre for Infectious Diseases, University of Heidelberg, Heidelberg, Germany
[5]FIND, Geneva, Switzerland
[6]Honorary Research Fellow, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom

Corresponding author
Mikashmi Kohli, PhD
Department of Epidemiology, Biostatistics and Occupational Health,
McGill University
Montreal, Canada
mikashmikohli@gmail.com

BACKGROUND
Globally, extrapulmonary tuberculosis accounted for 15% of the 7.0 million cases of tuberculosis notified in 2018, range, 8% in the WHO Western Pacific Region to 24% in the Eastern Mediterranean Region (WHO Global Tuberculosis Report 2019). In 2018, there were about 500,000 new cases of rifampicin-resistant tuberculosis; of these cases, 78% had multidrug-resistant tuberculosis (WHO Global Tuberculosis Report 2019). Xpert MTB/RIF and Xpert Ultra are WHO-recommended rapid tests that simultaneously detect tuberculosis and rifampicin resistance in people with signs and symptoms of tuberculosis and are suitable for use at lower levels of the health system. The aim of this systematic review was to determine the diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra in non-respiratory specimens for different forms of extrapulmonary tuberculosis in adults.

METHODS
Search methods
We searched the following databases: Cochrane Infectious Diseases Group Specialized Register; MEDLINE (OVID, from 1966); Embase (OVID, from 1974); Science Citation Index - Expanded (from 1900), Conference Proceedings Citation Index - Science (CPCI-S, from 1990), and BIOSIS Previews (from 1926), all three from the Web of Science; Scopus (Elsevier, from 1970); and Latin American Caribbean Health Sciences Literature (LILACS) (BIREME, from 1982). We also searched ClinicalTrials.gov, the WHO International Clinical Trials Registry (ICTRP) Platform (www.who.int/trialsearch), and the International Standard Randomized Controlled Trials Number (ISRCTN) registry (www.isrctn.com/) for trials in progress, and ProQuest Dissertations & Theses A&I (from 1990) for dissertations to 2 August 2019, without language restriction.

Selection criteria
We included cross-sectional and cohort studies using non-respiratory specimens that evaluated Xpert MTB/RIF, Xpert Ultra, or both against a microbiological and a composite reference standard for tuberculosis and culture-based drug susceptibility testing or MTBDR*plus* for rifampicin resistance. We included the following common forms of extrapulmonary TB: TB meningitis and pleural, lymph node, bone or joint, genitourinary, peritoneal, pericardial, and disseminated TB. We excluded studies that evaluated Xpert MTB/RIF or Xpert Ultra in gastric fluid, as this specimen is used most often to investigate pulmonary TB in children. We also excluded stool specimens because tuberculosis bacteria may be swallowed and passed into stool as a marker of pulmonary tuberculosis.

Data collection and analysis
Two review authors independently extracted data using a standardized form. We assessed study quality using QUADAS-2. Whenever possible, we extracted data on per participant rather than per specimen. For most studies, the number of specimens was the same as the number of participants. We performed meta-analyses to estimate pooled sensitivity and specificity separately for Xpert MTB/RIF and Xpert Ultra separately for the different forms of extrapulmonary tuberculosis (and the related specimens used for diagnosis), and rifampicin resistance. We used a bivariate random-effects model to determine summary estimates for sensitivity and specificity. We performed analyses by type of reference standard, microbiological or composite.

RESULTS
Detection of extrapulmonary tuberculosis
For detection of extrapulmonary tuberculosis, we included 65 studies, Figure 1. A total of 63 studies (13,144 participants) evaluated Xpert MTB/RIF and six studies (507 participants)

evaluated Xpert Ultra, including five studies that evaluated Xpert MTB/RIF and Xpert Ultra, Appendix.

Of the total 65 studies, 39 studies (60%) took place in high tuberculosis burden and 41 (63%) in high TB/HIV burden countries. We judged risk of bias to be low in the patient selection, index test, and flow and timing domains and high or unclear in the reference standard domain because many studies decontaminated sterile specimens before culture inoculation. Regarding applicability, in the patient selection domain, we judged high or unclear concern for most studies because either the participants were evaluated exclusively as inpatients at tertiary care centers (for any form of extrapulmonary TB, other than TB meningitis), or we were not sure about the clinical settings.

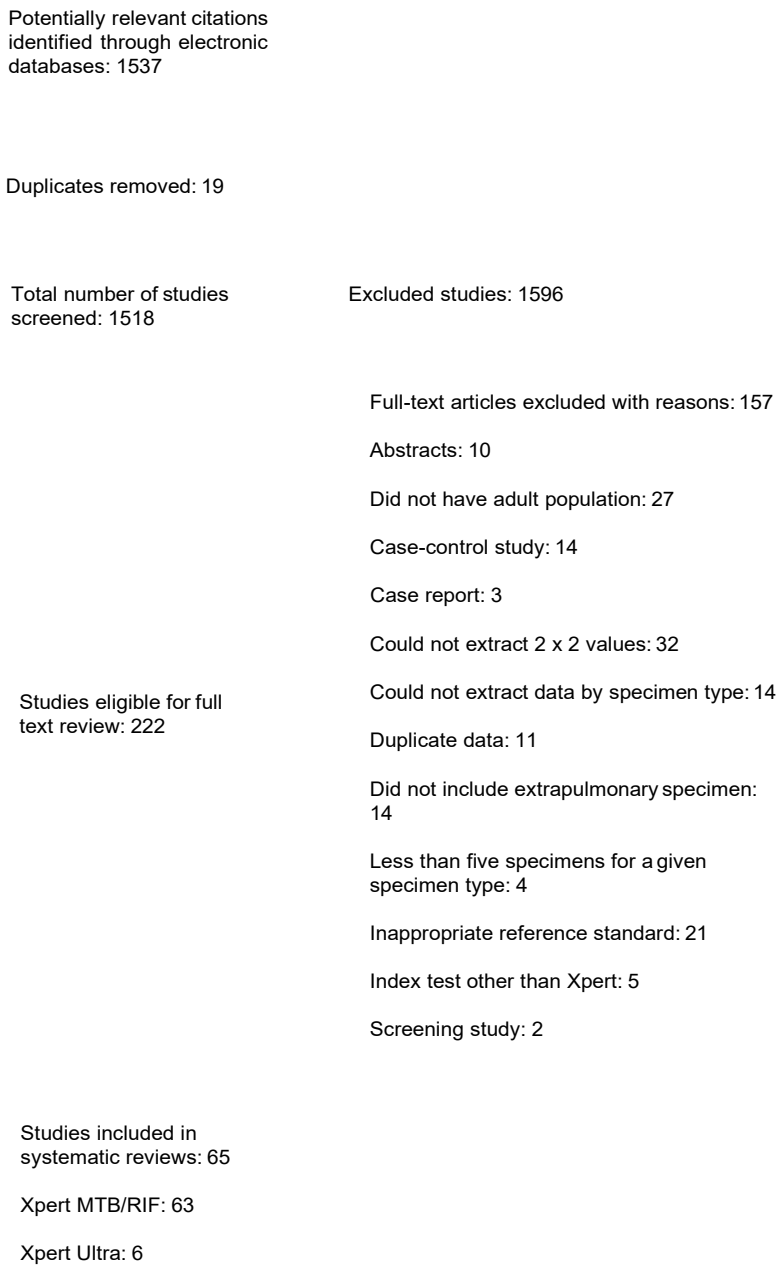


Figure 1. Flow of studies in the reivew

PICO 3, extrapulmonary TB in adults

32. Should Xpert MTB/RIF be used to diagnose TB meningitis in cerebrospinal fluid in adults with signs and symptoms of TB meningitis, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity (95% credible Interval (CrI)) were 70.3% (60.9 to 79.0) and 96.8% (95.2 to 98.1), (28 studies, 3103 participants; moderate-certainty evidence for sensitivity; high-certainty evidence for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 99 would be Xpert MTB/RIF-positive and 29 (29%) would not have tuberculosis (false-positives); 901 would be Xpert MTB/RIF-negative and 30 (3%) would have tuberculosis (false-negatives).

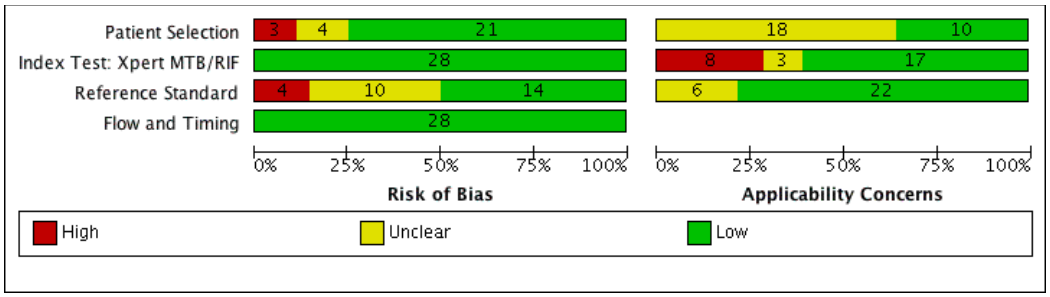


Figure 2. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

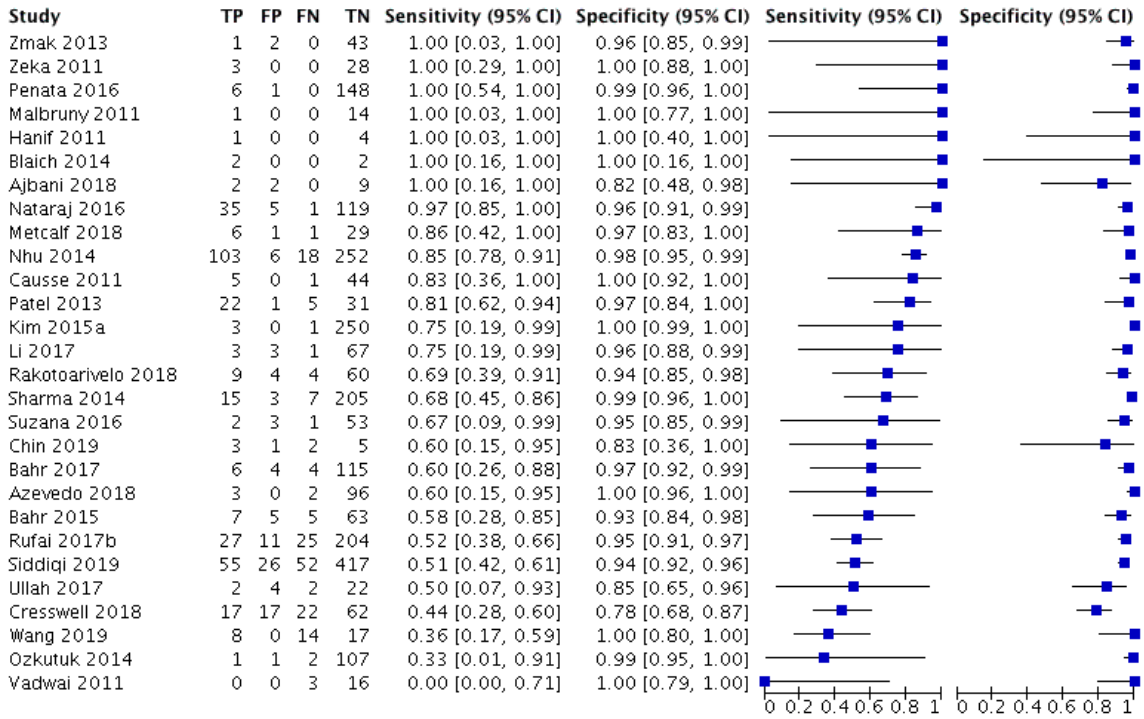


Figure 3. Xpert MTB/RIF for TB meningitis, against a microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line.

33. Should Xpert MTB/RIF be used to diagnose TB meningitis in cerebrospinal fluid in adults with signs and symptoms of TB meningitis, against a composite reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 40.6% (30.0 to 52.6) and 99.5% (98.9 to 99.9), (12 studies, 1897 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).
For a population of 1000 people where 100 have tuberculosis on culture, 45 would be Xpert MTB/RIF-positive and 4 (9%) would not have tuberculosis (false-positives); 955 would be Xpert MTB/RIF-negative and 59 (6%) would have tuberculosis (false-negatives).

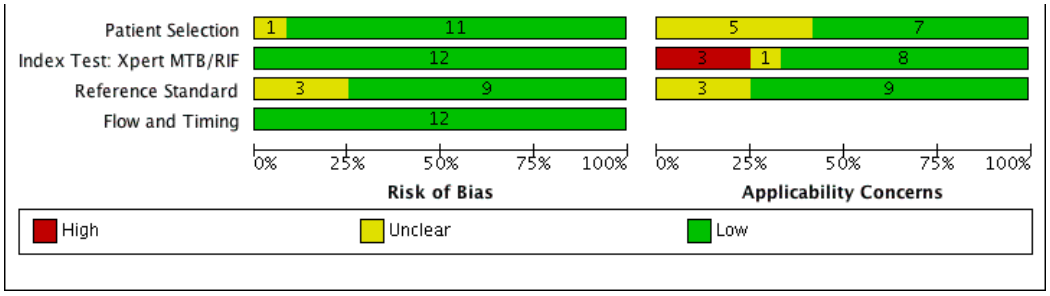


Figure 4. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

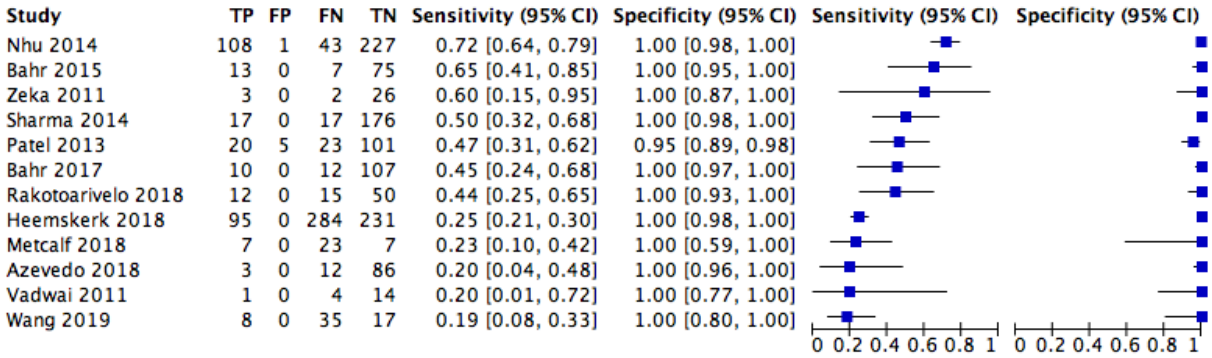


Figure 5. Xpert MTB/RIF for TB meningitis, against a composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

34. Should Xpert Ultra be used to diagnose TB meningitis in cerebrospinal fluid in adults with signs and symptoms of TB meningitis, against a microbiological reference standard?
Xpert Ultra pooled sensitivity and specificity were 86.9% (69.4 to 95.7) and 87.7% (69.0 to 95.6), (4 studies, 183 participants; low-certainty evidence).
The four studies include Chin 2019 (11 participants), which had a specificity of 50%. We could only explain in part the low specificity. Therefore, we conducted a sensitivity analysis removing Chin 2019. The pooled sensitivity and specificity of this analysis were 88.1% (69.4 to 96.5) and 91.9% (78.4 to 97.5), respectively (3 studies, 172 participants). When we removed Chin 2019 from the analysis, we noted that the pooled specificity increased to 91.9% as compared to the specificity of 87.7% when all four studies were included.

For a population of 1000 people where 100 have tuberculosis on culture, 210 would be Xpert Ultra-positive and 126 (60%) would not have tuberculosis (false-positives); 790 would be Xpert Ultra-negative and 16 (2%) would have tuberculosis (false-negatives).

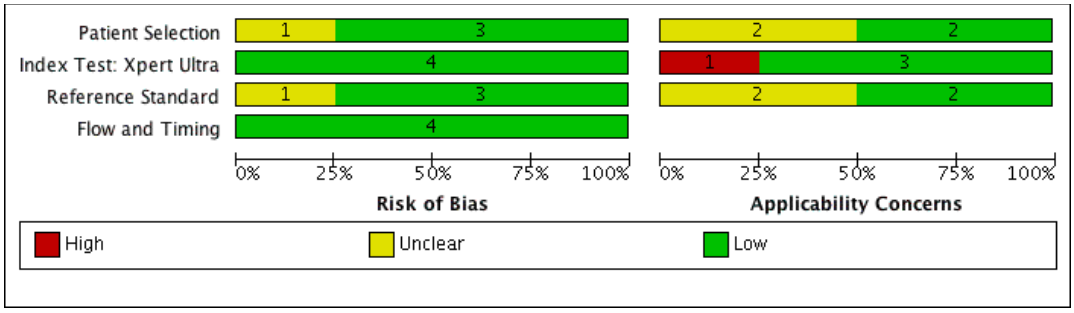


Figure 6. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

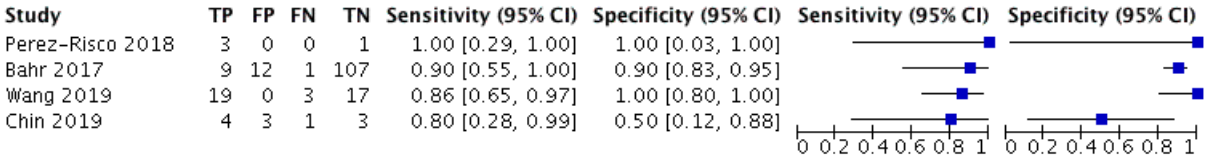


Figure 7. Xpert Ultra for TB meningitis, against a microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

35. Should Xpert Ultra be used to diagnose TB meningitis in CSF in adults with signs and symptoms of TB meningitis, against a composite reference standard?
Xpert Ultra sensitivity ranged from 44% to 70% and specificity ranged from 95% to 100%, (2 studies, 189 participants; very low-certainty evidence for sensitivity; very low-certainty evidence for specificity).

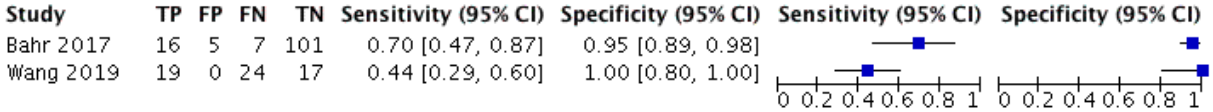


Figure 8. Xpert Ultra for TB meningitis, against a composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

36. Should Xpert MTB/RIF be used to diagnose lymph node TB in lymph node fluid in adults with signs and symptoms of lymph node TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 88.7% (82.3 to 93.2) and 86.0% (77.7 to 92.1), (14 studies, 1588 participants; moderate-certainty evidence for sensitivity; very low-certainty evidence for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 215 would be Xpert MTB/RIF-positive and 126 (59%) would not have tuberculosis (false-positives); 785 would be Xpert MTB/RIF-negative and 11(1%) would have tuberculosis (false-negatives).

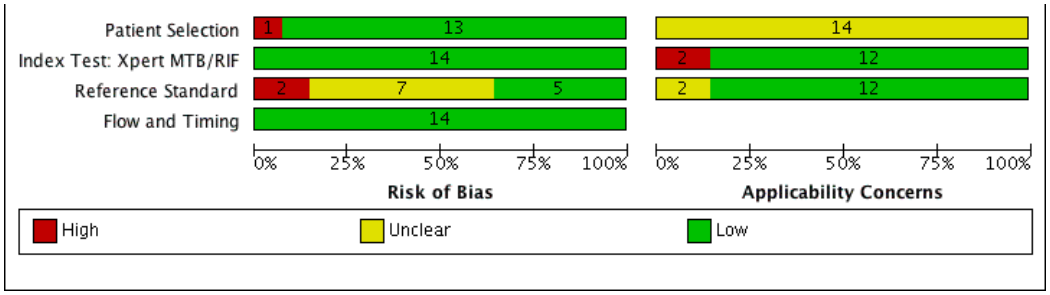
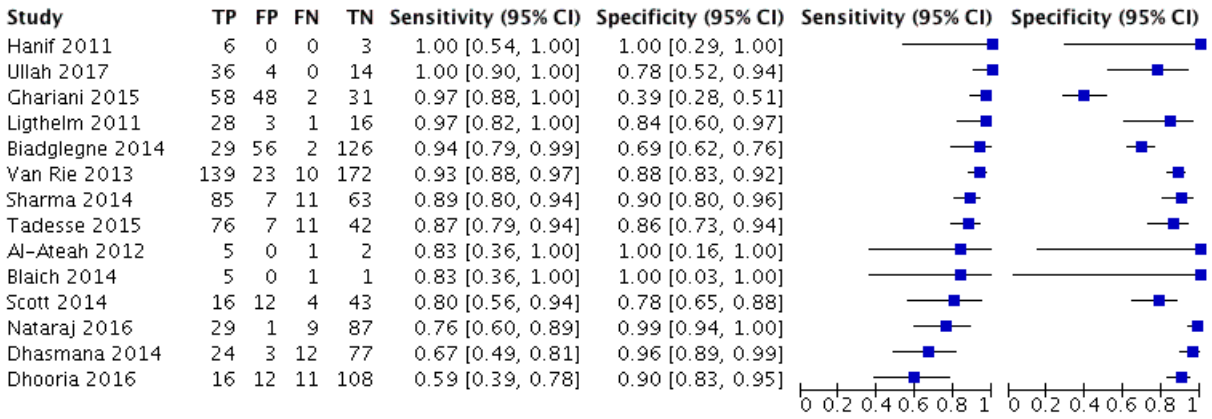


Figure 9. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.



Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Hanif 2011	6	0	0	3	1.00 [0.54, 1.00]	1.00 [0.29, 1.00]
Ullah 2017	36	4	0	14	1.00 [0.90, 1.00]	0.78 [0.52, 0.94]
Ghariani 2015	58	48	2	31	0.97 [0.88, 1.00]	0.39 [0.28, 0.51]
Ligthelm 2011	28	3	1	16	0.97 [0.82, 1.00]	0.84 [0.60, 0.97]
Biadglegne 2014	29	56	2	126	0.94 [0.79, 0.99]	0.69 [0.62, 0.76]
Van Rie 2013	139	23	10	172	0.93 [0.88, 0.97]	0.88 [0.83, 0.92]
Sharma 2014	85	7	11	63	0.89 [0.80, 0.94]	0.90 [0.80, 0.96]
Tadesse 2015	76	7	11	42	0.87 [0.79, 0.94]	0.86 [0.73, 0.94]
Al-Ateah 2012	5	0	1	2	0.83 [0.36, 1.00]	1.00 [0.16, 1.00]
Blaich 2014	5	0	1	1	0.83 [0.36, 1.00]	1.00 [0.03, 1.00]
Scott 2014	16	12	4	43	0.80 [0.56, 0.94]	0.78 [0.65, 0.88]
Nataraj 2016	29	1	9	87	0.76 [0.60, 0.89]	0.99 [0.94, 1.00]
Dhasmana 2014	24	3	12	77	0.67 [0.49, 0.81]	0.96 [0.89, 0.99]
Dhooria 2016	16	12	11	108	0.59 [0.39, 0.78]	0.90 [0.83, 0.95]

Figure 10. Xpert MTB/RIF for lymph node TB, against a microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

37. Should Xpert MTB/RIF be used to diagnose lymph node TB in lymph node fluid in adults with signs and symptoms of lymph node TB, against a composite reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 80.9% (62.1 to 92.0) and 95.9% (90.1 to 98.3), (4 studies, 679 participants; low-certainty evidence for sensitivity; low-certainty evidence for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 118 would be Xpert MTB/RIF-positive and 37 (31%) would not have tuberculosis (false-positives); 882 would be Xpert MTB/RIF-negative and 19(2%) would have tuberculosis (false-negatives).

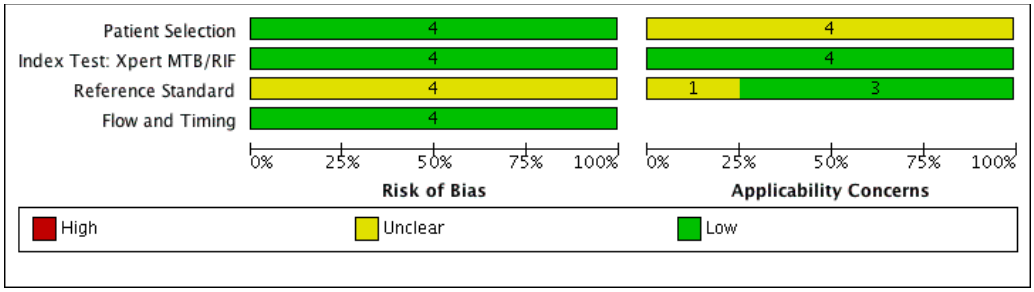


Figure 11. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

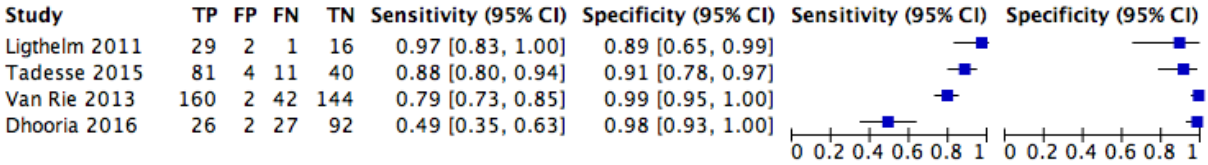


Figure 12. Xpert MTB/RIF for lymph node TB, against a composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

38. Should Xpert MTB/RIF be used to diagnose lymph node TB in lymph node tissue in adults with signs and symptoms of lymph node TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 82.0% (72.9 to 89.2) and 79.3% (58.5 to 90.6), (11 studies, 786 participants; moderate-certainty evidence for sensitivity; very low-certainty evidence for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 268 would be Xpert MTB/RIF-positive and 186 (65%) would not have tuberculosis (false-positives); 722 would be Xpert MTB/RIF-negative and 18 (2%) would have tuberculosis (false-negatives).

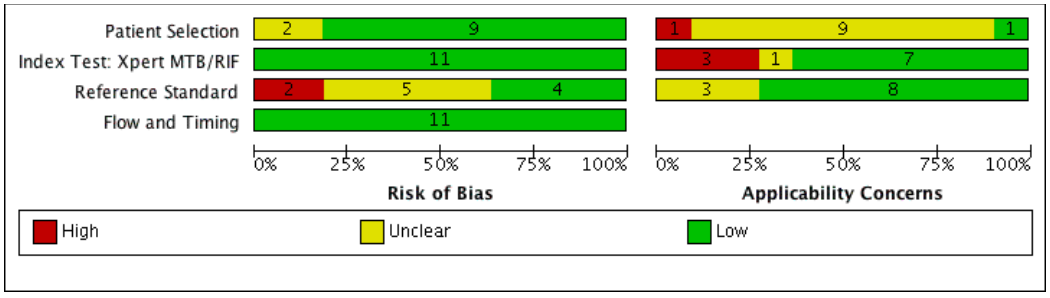


Figure 13. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

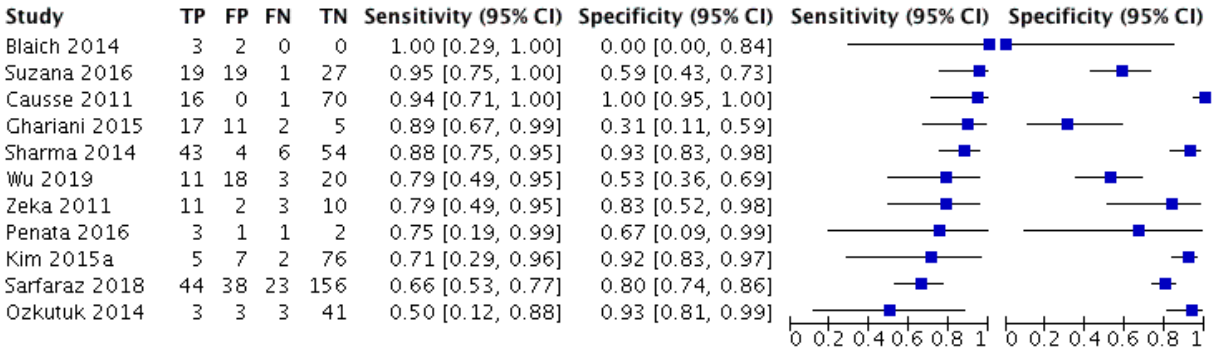


Figure 14. Xpert MTB/RIF for lymph node TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

39. Should Xpert Ultra be used to diagnose lymph node TB in lymph node tissue in adults with signs and symptoms of lymph node TB, against a microbiological reference standard?
Xpert Ultra reported a sensitivity of 100% (95 to 100) and specificity of 38% (22 to 55), (1 study, 50 participants; very low-certainty of evidence for sensitivity; very low- certainty of evidence for specificity).

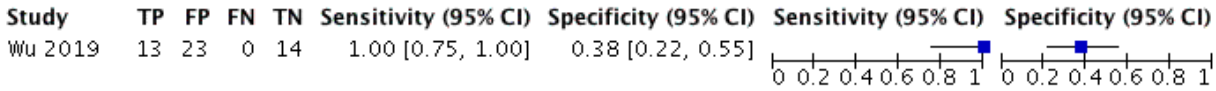


Figure 15. Xpert Ultra for lymph node TB, against microbiological reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

40. Should Xpert MTB/RIF be used to diagnose pleural TB in pleural fluid in adults with signs and symptoms of pleural TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 49.6% (39.3 to 60.5) and 98.7% (97.2 to 99.5), (24 studies, 2926 participants; very low-certainty evidence for sensitivity; high-certainty evidence for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 62 would be Xpert MTB/RIF-positive and 12 (19%) would not have tuberculosis (false-positives); 938 would be Xpert MTB/RIF-negative and 50 (5%) would have tuberculosis (false-negatives).

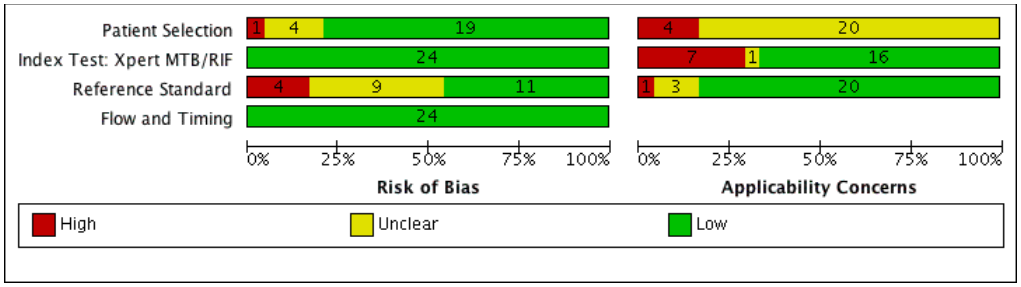


Figure 16. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

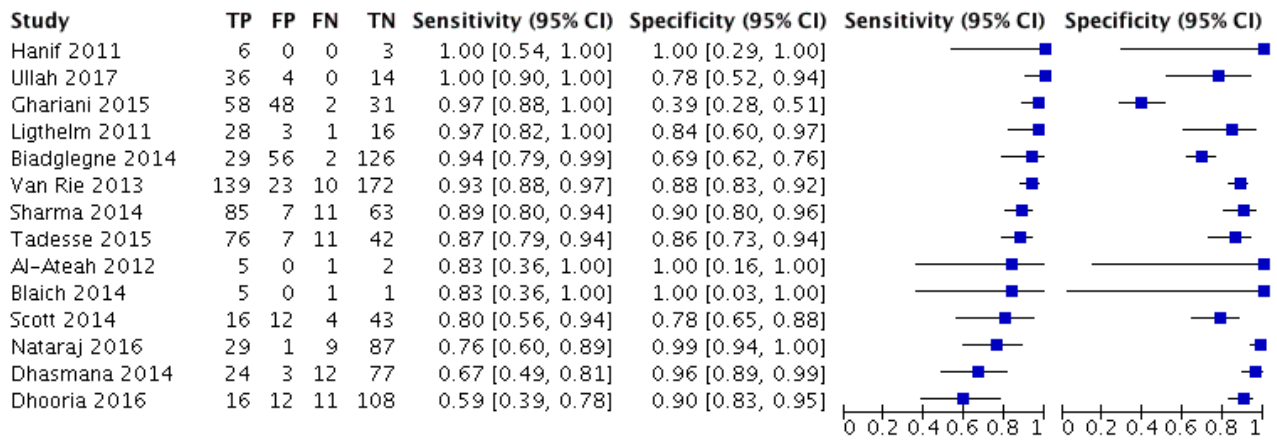


Figure 17. Xpert MTB/RIF for pleural TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

41. Should Xpert MTB/RIF be used to diagnose pleural TB in pleural fluid in adults with signs and symptoms of pleural TB, against a composite reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 19.3% (11.9 to 28.3) and 98.9% (97.5 to 99.6), (10 studies, 1024 participants; moderate-certainty evidence for sensitivity; high certainty for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 16 would be Xpert MTB/RIF-positive and 11 (69%) would not have tuberculosis (false-positives); 984 would be Xpert MTB/RIF-negative and 29 (2%) would have tuberculosis (false-negatives).

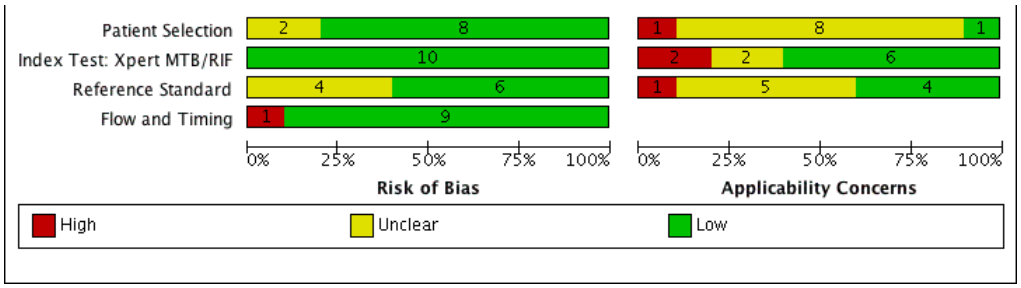


Figure 18. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

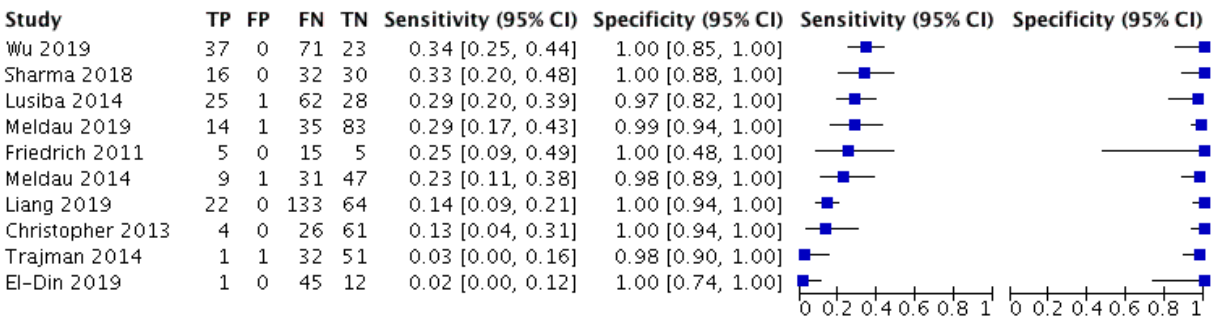


Figure 19. Xpert MTB/RIF for pleural TB, against composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

42. Should Xpert Ultra be used to diagnose pleural TB in pleural fluid in adults with signs and symptoms of pleural TB, against a microbiological reference standard?
Xpert Ultra pooled sensitivity and specificity were 71.1% (49.0 to 85.8) and 71.2% (52.3 to 85.5), (3 studies, 257 participants; very low-certainty evidence).

For a population of 1000 people where 100 have tuberculosis on culture, 330 would be Xpert Ultra-positive and 259 (78%) would not have tuberculosis (false-positives); 670 would be Xpert Ultra-negative and 29 (4%) would have tuberculosis (false-negatives).

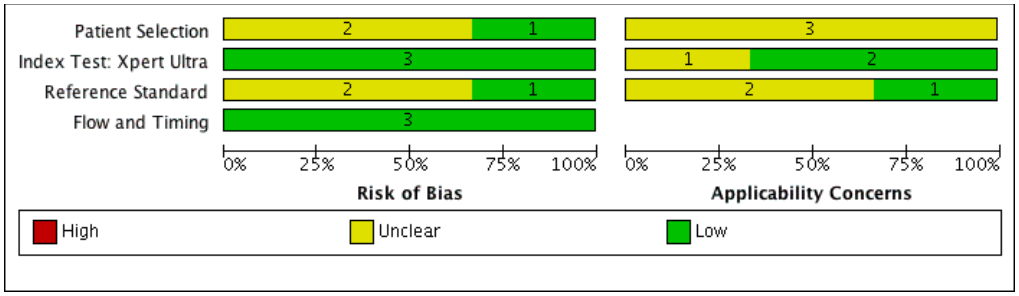


Figure 20. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

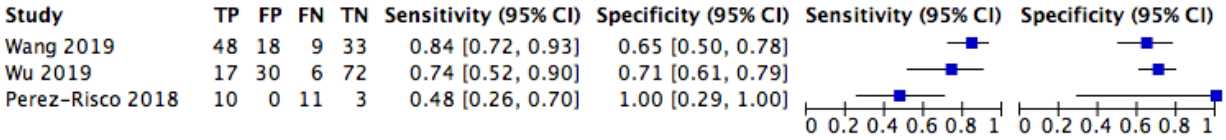


Figure 21. Xpert Ultra for pleural TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

43. Should Xpert Ultra be used to diagnose pleural TB in pleural fluid in adults with signs and symptoms of pleural TB, against a composite reference standard?

Xpert Ultra sensitivity ranged from 38% to 61% and specificity ranged from 96% to 99%, (2 studies, 263 participants; very low-certainty evidence for sensitivity; moderate certainty for specificity).

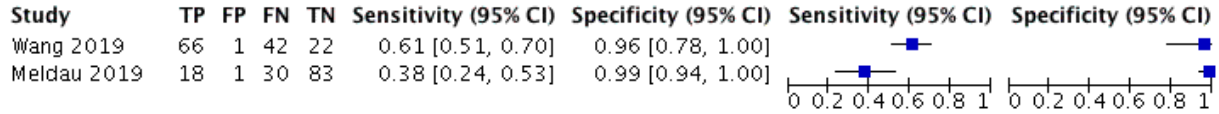


Figure 22. Xpert Ultra for pleural TB, against composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

44. Should Xpert MTB/RIF be used to diagnose genitourinary TB in urine in adults with signs and symptoms of genitourinary TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 84.7% (70.8 to 93.1) and 97.3% (91.0 to 99.2), (9 studies, 943 participants; very low-certainty evidence for sensitivity; moderate certainty for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 109 would be Xpert MTB/RIF-positive and 24 (22%) would not have tuberculosis (false-positives); 891 would be Xpert MTB/RIF -negative and 15 (2%) would have tuberculosis (false-negatives).

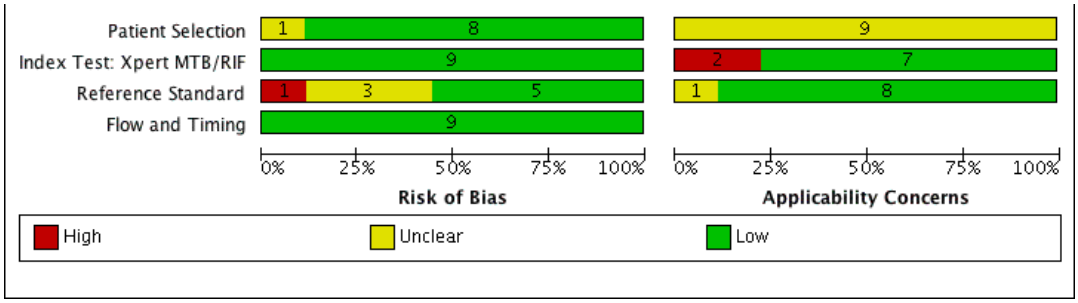


Figure 23. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

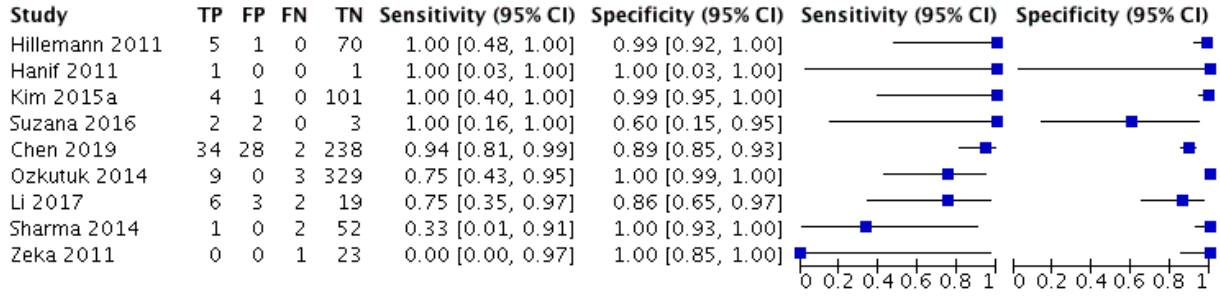


Figure 24. Xpert MTB/RIF for genitourinary TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and

specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

45. Should Xpert MTB/RIF be used to diagnose genitourinary TB in urine in adults with signs and symptoms of genitourinary TB, against a composite reference standard?
Xpert MTB/RIF sensitivity ranged from 33% to 41% and specificity was 100%, (2 studies, 463 participants; low certainty evidence)

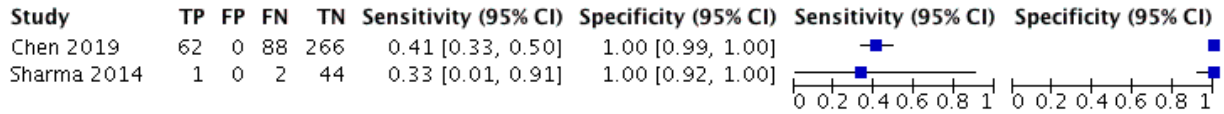


Figure 25. Xpert MTB/RIF for genitourinary TB, against composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

46. Should Xpert Ultra be used to diagnose genitourinary TB in urine in adults with signs and symptoms of genitourinary TB, against a microbiological reference standard?
Xpert Ultra reported a sensitivity and specificity of 100% (1 study, 24 participants; very low certainty evidence)

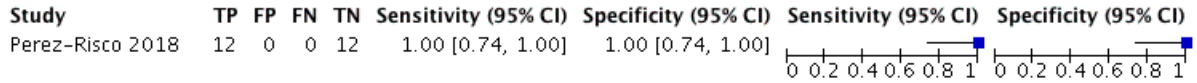


Figure 26. Xpert Ultra for genitourinary TB, against microbiological reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

47. Should Xpert MTB/RIF be used to diagnose bone or joint TB in bone or joint fluid in adults with signs and symptoms of bone or joint TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 97.1% (91.7 to 99.2) and 93.7% (66.7 to 99.1), (6 studies, 471 participants; moderate-certainty evidence for sensitivity; very low certainty for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 154 would be Xpert MTB/RIF-positive and 57 (37%) would not have tuberculosis (false-positives); 846 would be Xpert MTB/RIF -negative and 3 (0.3%) would have tuberculosis (false-negatives).

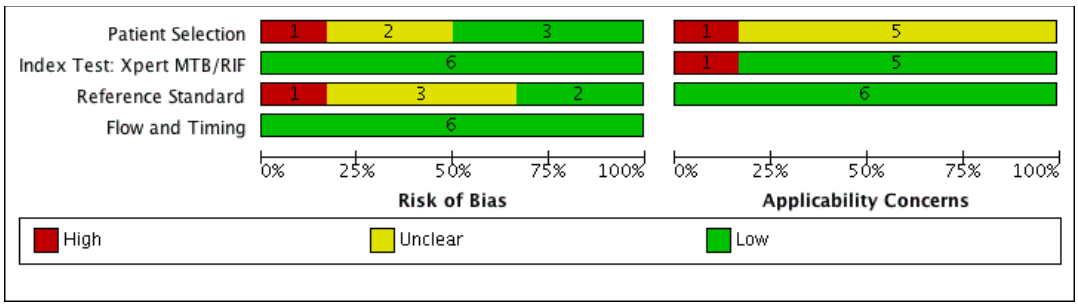


Figure 27. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

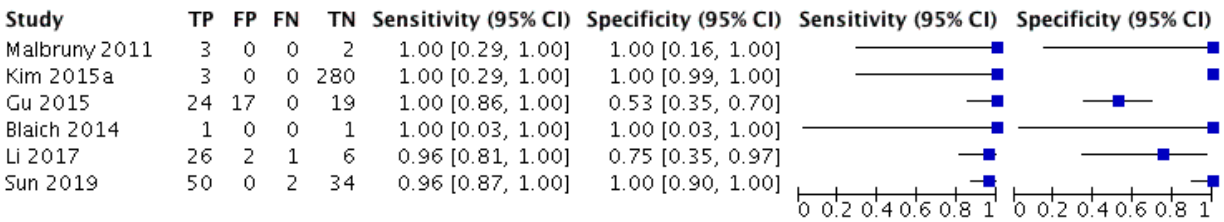


Figure 28. Xpert MTB/RIF for bone or joint TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

48. Should Xpert MTB/RIF be used to diagnose bone or joint TB in bone or joint fluid in adults with signs and symptoms of bone or joint TB, against a composite reference standard?
Xpert MTB/RIF sensitivity ranged from 82% to 94% and specificity was 100% (2 studies, 205 participants; low certainty for sensitivity; very low-certainty for specificity)

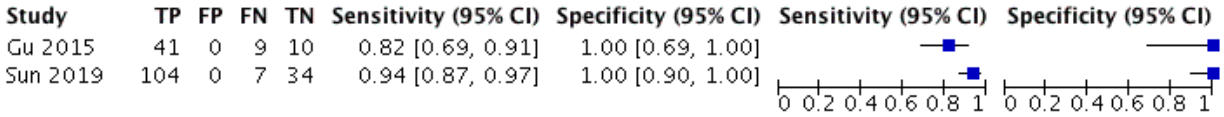


Figure 29. Xpert MTB/RIF for bone or joint TB, against composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

49. Should Xpert Ultra be used to diagnose bone or joint TB in bone or joint fluid in adults with signs and symptoms of bone or joint TB, against a microbiological reference standard?
Xpert Ultra sensitivity was 96% (87% to 100%) and specificity was 97% (85% to 100%), (1 study, 86 participants; very low-certainty evidence)

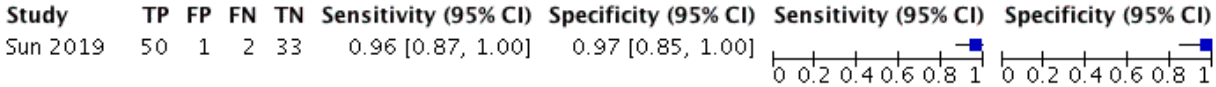


Figure 30. Xpert Ultra for bone or joint TB, against microbiological reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

50. Should Xpert Ultra be used to diagnose bone or joint TB in bone or joint fluid in adults with signs and symptoms of bone or joint TB, against a composite reference standard?
Xpert Ultra sensitivity was 96% (87% to 100%) and specificity was 97% (85% to 100%), (1 study, 86 participants; low-certainty evidence for sensitivity; very low- certainty for specificity)

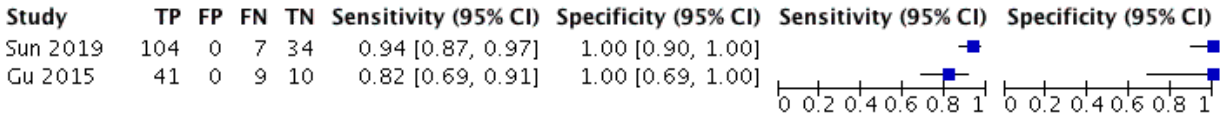


Figure 31. Xpert Ultra for bone or joint TB, against composite reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

51. Should Xpert MTB/RIF be used to diagnose peritoneal TB in peritoneal fluid in adults with signs and symptoms of peritoneal TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 58.9% (42.3 to 75.8) and 97.3% (95.1 to 98.7), (13 studies, 619 participants; low-certainty evidence for sensitivity; high certainty for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 83 would be Xpert MTB/RIF-positive and 24 (29%) would not have tuberculosis (false-positives); 917 would be Xpert MTB/RIF -negative and 41 (5%) would have tuberculosis (false-negatives).

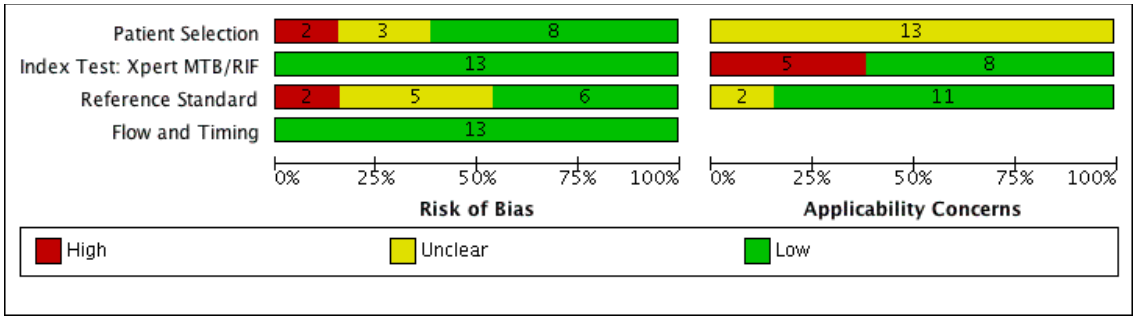


Figure 32. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

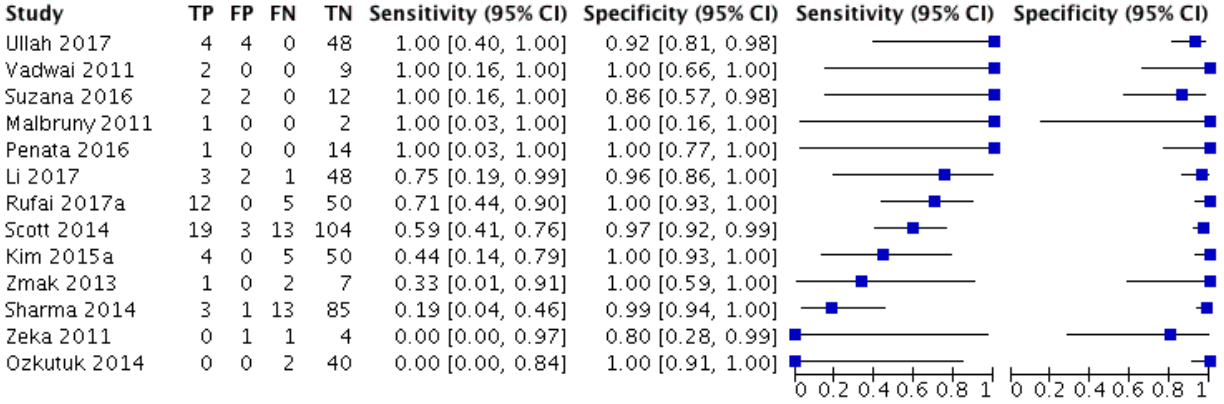


Figure 33. Xpert MTB/RIF for peritoneal TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

52. Should Xpert MTB/RIF be used to diagnose pericardial TB in pericardial fluid in people with signs and symptoms of pericardial TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 60.4% (34.7 to 81.7) and 87.8% (72.7 to 96.9), (5 studies, 181 participants; very low-certainty evidence for sensitivity; low certainty for specificity).

For a population of 1000 people where 100 have tuberculosis on culture, 170 would be Xpert MTB/RIF-positive and 110 (65%) would not have tuberculosis (false-positives); 830 would be Xpert MTB/RIF -negative and 40 (5%) would have tuberculosis (false-negatives).

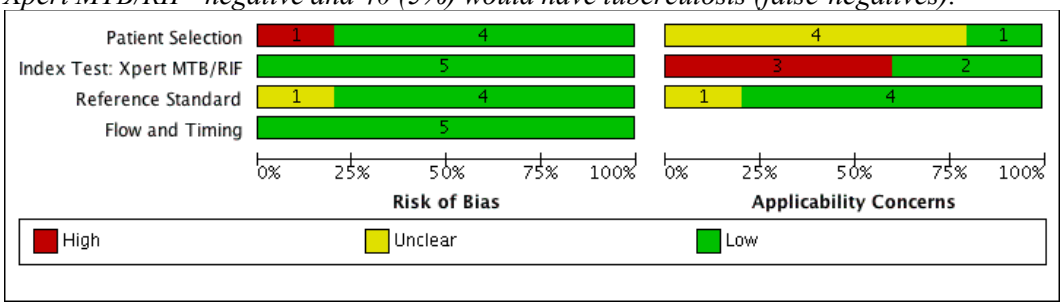


Figure 34. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

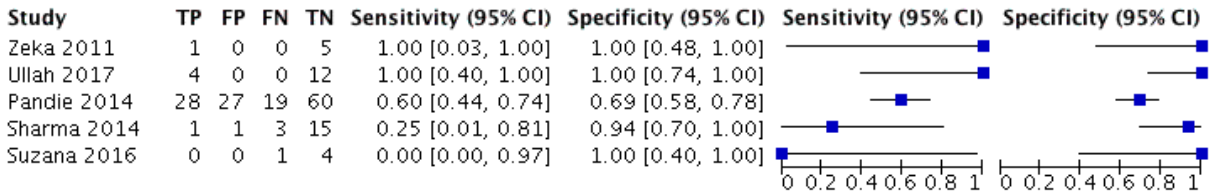


Figure 35. Xpert MTB/RIF for pericardial TB, against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

53. Should Xpert MTB/RIF be used to diagnose pericardial TB in pericardial fluid in people with signs and symptoms of pericardial TB, against a composite reference standard?
Xpert MTB/RIF sensitivity ranged from 40% to 75% and specificity ranged from 99% to 100%, (2 studies, 77 participants, very low- certainty of evidence)

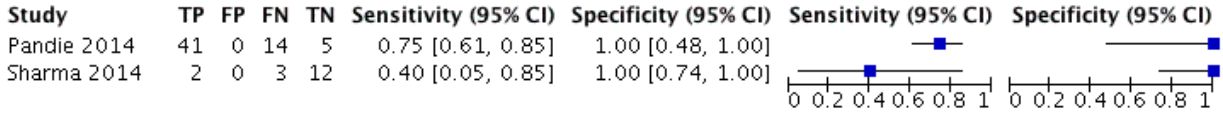


Figure 36. Xpert MTB/RIF for pericardial TB, against composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

54. Should Xpert MTB/RIF be used to diagnose disseminated TB in blood in people with signs and symptoms of disseminated TB, against a microbiological reference standard?
Xpert MTB/RIF reported a sensitivity of 56% (21% to 86%) and specificity of 94% (84% to 98%), (1 study, 74 participants, very low- certainty of evidence).

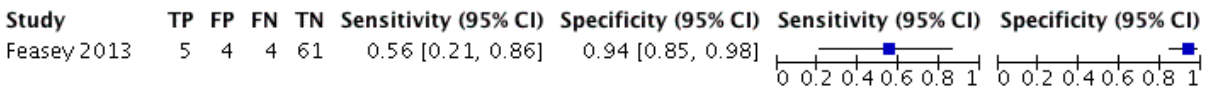


Figure 37. Xpert MTB/RIF for disseminated TB, against microbiological reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

Table. PICO 3, diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for extrapulmonary TB and rifampicin resistance in adults

PICO sub question	Test, Analysis Group	Number of studies (specimens)	Reference Standard	Pooled Sensitivity (95% CrI)	Pooled Specificity (95% CrI)	Positive Predictive Value (95% CrI)	Negative Predictive Value (95% CrI)
32	Xpert MTB/RIF, CSF	28 (3103)	MRS	70.3% (60.9 to 79.0)	96.8% (95.2 to 98.1)	71% (59 to 82)	97% (96 to 98)
33	Xpert MTB/RIF, CSF	12 (1897)	CRS	40.6% (30.0 to 52.6)	99.5% (98.9 to 99.9)	91% (75 to 98)	94% (93 to 95)
34	Xpert Ultra, CSF	4 (184)	MRS	86.9% (69.4 to 95.7)	87.7% (69.0 to 95.6)	40% (17 to 68)	98% (94 to 99)
35	Xpert Ultra, CSF	2 (189)	CRS	-	-	-	-
36	Xpert MTB/RIF, lymph node fluid	14 (1588)	MRS	88.7% (82.3 to 93.2)	86.0% (77.7 to 92.1)	41% (29 to 57)	99% (97 to 99)
37	Xpert MTB/RIF, lymph node fluid	4 (679)	CRS	80.9% (62.1 to 92.0)	95.9% (90.1 to 98.3)	69% (41 to 86)	98% (96 to 99)
38	Xpert MTB/RIF, lymph node tissue	11 (786)	MRS	82.0% (72.9 to 89.2)	79.3% (58.5 to 90.6)	31% (16 to 51)	98% (95 to 99)
39	Xpert Ultra, lymph node tissue	1 (50)	MRS	-	-	-	-
40	Xpert MTB/RIF, pleural fluid	24 (2926)	MRS	49.6% (39.3 to 60.5)	98.7% (97.2 to 99.5)	81% (61 to 94)	95% (93 to 96)
41	Xpert MTB/RIF, pleural fluid	10 (1024)	CRS	19.3% (11.9 to 28.3)	98.9% (97.5 to 99.6)	66% (35 to 88)	92% (91 to 97)

42	Xpert Ultra, pleural fluid	3 (257)	MRS	71.1% (49.0 to 85.8)	71.2% (52.3 to 85.5)	22% (10 to 40)	96% (90 to 98)
43	Xpert Ultra, pleural fluid	2 (263)	CRS	-	-	-	-
44	Xpert MTB/RIF, urine	9 (943)	MRS	84.7% (70.8 to 93.1)	97.3% (91.0 to 99.2)	93% (77 to 98)	94% (88 to 97)
45	Xpert MTB/RIF, urine	2 (463)	CRS	-	-	-	-
46	Xpert Ultra ,	1 (24)	MRS	-	-	-	-
47	Xpert MTB/RIF, bone or joint fluid	6 (471)	MRS	97.1% (91.7 to 99.2)	93.7% (66.7 to 99.1)	63% (23 to 93)	100% (99 to 100)
48	Xpert MTB/RIF, bone or joint fluid	2 (161)	CRS	-	-	-	-
49	Xpert Ultra, bone or joint fluid	2 (94)	MRS	-	-	-	-
50	Xpert Ultra, bone or joint fluid	1 (145)	CRS	-	-	-	-
51	Xpert MTB/RIF, peritoneal fluid	13 (580)	MRS	58.9% (42.3 to 75.8)	97.3% (95.1 to 98.7)	71% (49 to 86)	96% (94 to 97)
52	Xpert MTB/RIF, pericardia l fluid	5 (181)	MRS	60.4% (34.7 to 81.7)	87.8% (72.7 to 96.9)	35% (12 to 75)	95% (91 to 98)
53	Xpert MTB/RIF, pericardia l fluid	2 (77)	CRS	-	-	-	-
54	Xpert MTB/RIF, blood	2 (85)	MRS	-	-	-	-

55	Xpert MTB/RIF, rifampicin resistance	23 (1084)	Culture-based DST, LPA	96.3% (92.1 to 98.6)	98.8% (97.9 to 99.5)	91% (83 to 96)	100% (99 to 100)
56	Xpert Ultra, rifampicin resistance	3 (103)	Culture-based DST, LPA	96.7% (81.6 to 99.8)	98.8% (94.3 to 99.9)	90% (62 to 99)	100% (98 to 100)

Predictive values were determined for a pre-test probability of 10%; dashes indicate that there were insufficient data to perform a meta-analysis. Abbreviations: CrI: credible interval; CRS: composite: DST: drug susceptibility testing; LPA: line probe assay; MRS: microbiological

Detection of rifampicin resistance

For detection of rifampicin resistance, we included 35 studies. A total of 32 studies (1220 participants) evaluated Xpert MTB/RIF alone and three studies (79 participants) evaluated Xpert Ultra.

Of the total 35 studies, 17 took place in high MDR-TB burden countries. Most studies had low risk of bias and low concern about applicability.

55. Should Xpert MTB/RIF in extrapulmonary specimens be used to diagnose rifampicin resistance in adults with presumed extrapulmonary tuberculosis?

Xpert MTB/RIF pooled sensitivity and specificity were 96.3% (92.1 to 98.6) and 98.8% (97.9 to 99.5), (23 studies, 1084 patients; high-certainty evidence).

For a population of 1000 people where 100 have rifampicin-resistant TB, 106 would be positive for rifampicin-resistant TB: of these, 10 (9%) would not have rifampicin resistance (false-positives); and 894 would be negative for rifampicin-resistant TB: of these, 4 (0.4%) would have rifampicin resistance (false-negatives).

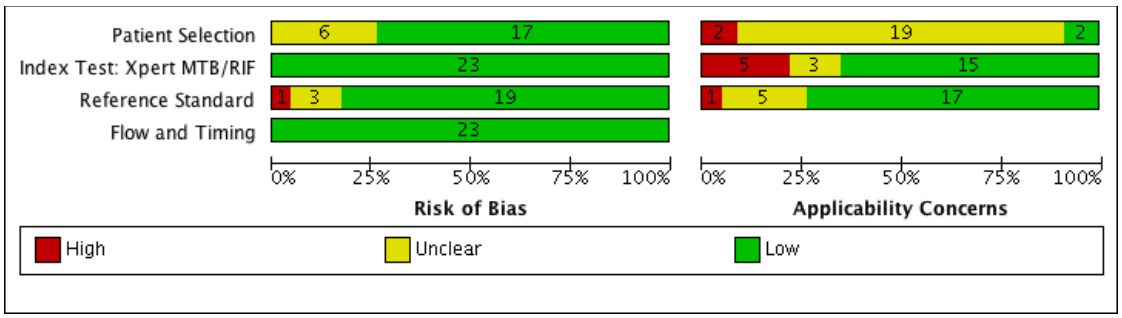


Figure 38. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

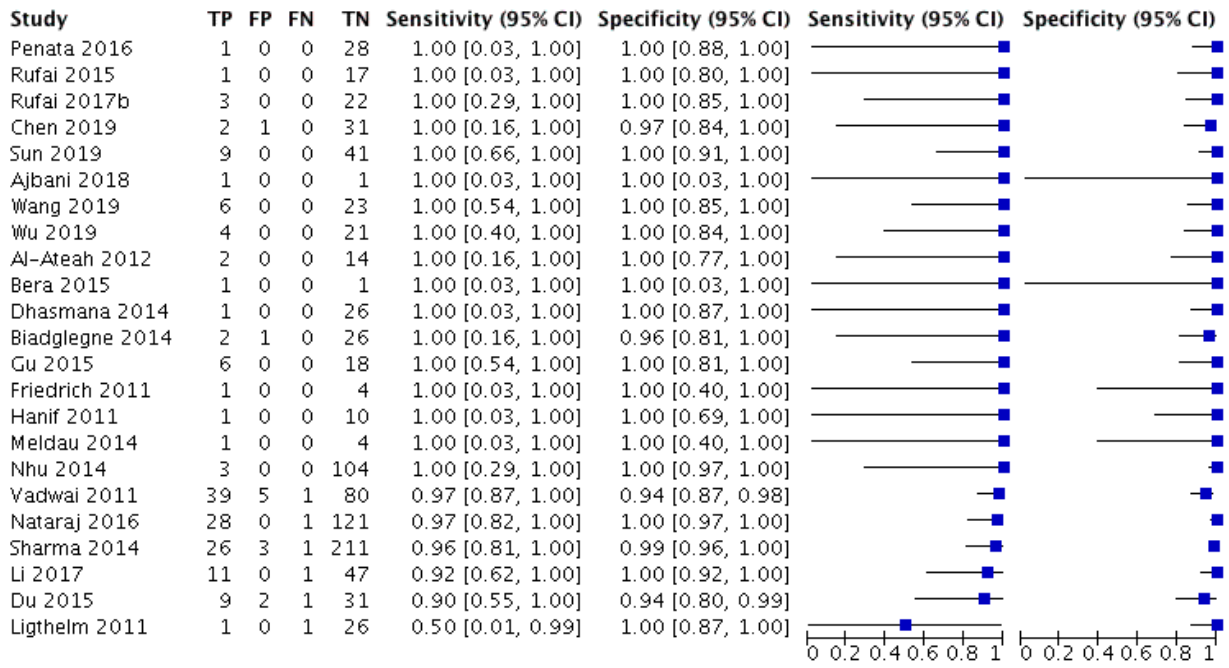


Figure 39. Xpert MTB/RIF for rifampicin resistance against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

56. Should Xpert Ultra in extrapulmonary specimens be used to diagnose rifampicin resistance in adults with presumed extrapulmonary tuberculosis?
Xpert MTB/RIF pooled sensitivity and specificity were 96.7% (81.6 to 99.8) and 98.8% (94.3 to 99.9), (3 studies, 103 patients; low-certainty evidence).

For a population of 1000 people where 100 have rifampicin-resistant TB, 108 would be positive for rifampicin-resistant TB: of these, 10 (10%) would not have rifampicin resistance (false-positives); and 892 would be negative for rifampicin-resistant TB: of these, 3 (0.3%) would have rifampicin resistance (false-negatives).

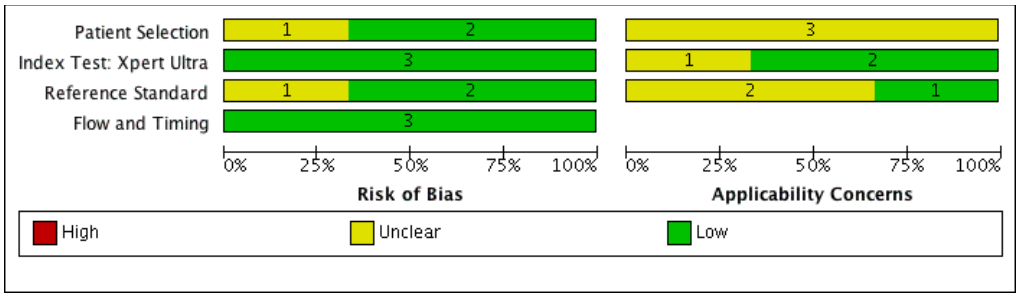


Figure 40. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

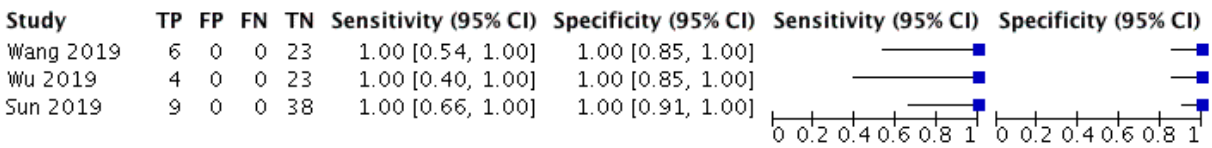


Figure 41. Xpert Ultra for rifampicin resistance against microbiological reference standard. The individual studies are ordered by decreasing sensitivity. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. FN: false-negative; FP: false-positive; TN: true-negative; TP: true-positive.

AUTHORS' CONCLUSIONS

Xpert MTB/RIF pooled sensitivity (defined by culture) varied across different types of specimens (50% in pleural fluid to 97% in bone or joint fluid). Xpert MTB/RIF pooled sensitivity was 80% or greater in lymph node fluid, lymph node tissue, urine, and bone or joint fluid. Xpert MTB/RIF pooled specificity (defined by culture) varied less than sensitivity across different specimens (79% in lymph node tissue to 99% in pleural fluid). Xpert MTB/RIF pooled specificity was 96% or greater in cerebrospinal fluid, pleural fluid, urine, and peritoneal fluid. For TB meningitis, Xpert Ultra had higher pooled sensitivity (84%) than Xpert MTB/RIF (70%) and lower pooled specificity (88%) than Xpert MTB/RIF (97%) based on an indirect comparison. Xpert MTB/RIF and Xpert Ultra had similar sensitivity and specificity for rifampicin resistance.

IMPLICATIONS FOR RESEARCH

Future studies should perform comparisons of different tests, including Xpert Ultra, as this approach will reveal which tests (or strategies) yield superior diagnostic accuracy. For these studies, the preferred study design is one in which all participants receive all available diagnostic tests or are randomly assigned to receive one or another of the tests. Studies should include children and HIV-positive people. Future research should acknowledge the concern associated with culture as a reference standard in paucibacillary specimens and should consider ways to address this limitation.

Rapid point-of-care diagnostic tests for extrapulmonary TB are critically needed. Research groups should focus on developing diagnostic tests and strategies that use readily available clinical specimens such as urine, rather than specimens that require invasive procedures for collection. Operational research is needed to ensure tests are optimally used in settings of intended use.

ACKNOWLEDGEMENTS

The editorial base of the Cochrane Infectious Diseases Group (CIDG) is funded by UK aid from the UK government for the benefit of low- and middle-income countries (project number 300342-104). The views expressed do not necessarily reflect the UK government's official policies.

We thank all authors of the included studies for answering our questions and providing additional data. We are grateful to Vittoria Lutje, CIDG Information Specialist, for help with the search strategy.

We thank Hannah Ryan, from the Tropical and Infectious Diseases Unit, Royal Liverpool University Hospital, for her help with the protocol. We thank Chunli Lu, from the Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing and Marcela Perlwitz, Ivy Tech Community College of Indiana, West Lafayette, for help with translation. We thank all authors of the included studies for providing answers to our questions along with additional data. We thank the referees for their helpful comments.

CONTRIBUTIONS OF AUTHORS

MK and KRS wrote early drafts of the protocol. CMD and SGS contributed methodological advice. KD contributed clinical expertise to the Cochrane Review. CMD and SGS tailored QUADAS-2 to the review. MK and KRS reviewed the studies and extracted accuracy data. MK and KRS assessed the methodological quality of included studies. IS and ND performed the statistical analyses. All review authors interpreted the findings. MK, ND, and KRS wrote the first draft of the review. MK and KRS prepared the 'Summary of findings' tables. All review authors contributed to the final manuscript.

DECLARATIONS OF INTEREST

Development of the systematic review was in part made possible with financial support from the United States Agency for International Development (USAID) administered by the World Health Organization (WHO) Global TB Programme, Switzerland. KRS, MK and IS received funding to carry out the review from USAID. In addition, KRS received financial support for the submitted work from the CIDG, and has received financial support for the preparation of systematic reviews and educational materials, consultancy fees from FIND (for the preparation of a systematic review), honoraria, and travel support to attend WHO guideline meetings. CMD and SGS work for FIND. FIND is a non-for-profit foundation whose mission is to find diagnostic solutions to overcome diseases of poverty in low- and middle-income countries. FIND works closely with the private and public sectors and receives funding from donors and some of its industry partners. FIND has an independent Scientific Advisory Committee and organizational firewalls that protect it against any undue influences in its work or in publication of its findings. More information on FIND's policy and guidelines for working with private sector partners can be found at www.finddx.org/business-model. ND received funding from the CIDG. We have no financial involvement with any organization or entity that has a financial interest in, or financial conflict with, the subject matter or materials discussed in the review apart from those disclosed.

Web Annex D.4. Xpert MTB/RIF and Xpert Ultra for detecting active tuberculosis in children: an updated systematic review

Alexander W Kay[1], Lucia González Fernández[1], Yemisi Takwoingi[2], Michael Eisenhut[3], Ryan D Vu[1], Karen R Steingart[4], Anne K Detjen[5], Gemma Villanueva[6], Hanna Bergman[6], Anna M Mandalakas[1]

[1]Global Tuberculosis Program, Baylor College of Medicine, Houston, Texas, USA
[2]Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom
[3]Paediatric Department, Luton & Dunstable University Hospital NHS Foundation Trust, Luton, United Kingdom
[4]Honorary Research Fellow, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom
[5]Health Section, United Nations Children's Fund (UNICEF), New York, New York, USA
[6]Cochrane Response, Cochrane, London, United Kingdom

Corresponding author:
Alexander W Kay, MD
Pediatrics, Baylor College of Medicine
Houston, Texas
USA
alexander.kay@bcm.edu

BACKGROUND

Globally, child tuberculosis accounted for 11% of the 10.0 million estimated global cases of tuberculosis in 2018. Children account for a disproportionate share of tuberculosis mortality (14%) suggesting poorer access to diagnosis and treatment. In 2018, there were about half a million new cases of rifampicin-resistant tuberculosis, and of these, 78% had multidrug-resistant tuberculosis; however programmatic data on the burden of drug-resistant tuberculosis in children are limited (WHO Global Tuberculosis Report 2019). Xpert MTB/RIF and Xpert Ultra are WHO-recommended rapid tests that simultaneously detect tuberculosis and rifampicin resistance in people with signs and symptoms of tuberculosis. The aim of this systematic review was to determine the diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for pulmonary tuberculosis in children with signs and symptoms of pulmonary tuberculosis, in several types of clinical specimens, including sputum specimens, gastric specimens, nasopharyngeal specimens, and stool specimens. In addition, this review aimed to determine the diagnostic accuracy of Xpert MTB/RIF for the diagnosis of lymph node tuberculosis and tuberculous meningitis in children.

In 2013, informed by a systematic review (Detjen and Mandalakas 2015), the WHO recommended the use of Xpert MTB/RIF in children as a front-line test for the diagnosis of tuberculosis. The review found that, when evaluated against a reference standard of culture, Xpert MTB/RIF had a sensitivity and specificity of 62% (95% credible interval (95%CrI) 51% to 73%) and 98% (95%CrI 97% to 99%) for the diagnosis of pulmonary TB in children. In preparation for a WHO meeting to update recommendations on the use of molecular methods for diagnosing tuberculosis, we performed a Cochrane Review to update the Detjen and Mandalakas review and to assess the accuracy of Xpert MTB/RIF and Xpert Ultra against both a microbiological reference standard and a composite reference standard. References of data included in this systematic review are included in Appendix A.

Clinical pathway

An example of the clinical pathway in children and the placement of the index tests within the diagnostic pathway can be found in Appendix B. A careful clinical history of tuberculosis exposure and symptoms is the first step in the diagnostic pathway for childhood tuberculosis. Children with household or other close and persistent exposure to a person with tuberculosis are at increased risk of tuberculosis infection and resultant progression to tuberculosis disease. All children with recent exposure to tuberculosis must be evaluated for clinical symptoms and examination findings consistent with tuberculosis disease. Additional testing depends on the context, but may include chest radiograph and a test of tuberculosis infection. Symptoms of tuberculosis disease are generally persistent for greater than two weeks and are unremitting (Marais 2005).The most common symptoms are cough, fever, decreased appetite, weight loss or failure to thrive, and fatigue or reduced playfulness. Symptoms of extrapulmonary tuberculosis are typically localized, and diagnostic findings are generally obtained from the site of disease (Appendix B). However, no symptom-based diagnostic algorithms have been validated or have been shown to be reliable in multiple contexts. Symptom-based diagnostic algorithms tend to perform poorly in children under three years of age and children living with HIV, two populations at high risk for disease progression (Marais 2006b).

Unfortunately, there are no examination features specific to pulmonary tuberculosis in children. However, the examination findings in extrapulmonary tuberculosis can be quite specific when identified. Clinicians should consider medical comorbidities that increase the risk for tuberculosis disease and modify diagnostic algorithms accordingly. HIV infection not only significantly increases risk of tuberculosis in the paediatric population, but also raises the risk of increased

disease severity. HIV-positive children often present with advanced forms of tuberculosis and have high levels of immunosuppression, further complicating diagnosis and management.

Additional diagnostic imaging studies can assist in the diagnosis of nearly all forms of pulmonary tuberculosis and extrapulmonary tuberculosis. Tests of tuberculosis infection, such as interferon gamma release assays or tuberculin skin tests, can also aid in establishing the diagnosis of tuberculosis in a child but are not necessary to make the diagnosis. Diagnostic recommendations strongly suggest collecting appropriate specimens from the suspected sites of involvement in both pulmonary and extrapulmonary tuberculosis for microbiological examination. The preferred sample in pulmonary tuberculosis is sputum, however in young children that cannot expectorate, the sample is commonly obtained via a gastric aspiration or sputum induction. To diagnose extrapulmonary tuberculosis, the collection of samples targets the affected site of disease.

The purpose of Xpert MTB/RIF or Xpert Ultra is diagnosis of active tuberculosis (pulmonary and extrapulmonary tuberculosis) and detection of rifampicin resistance. The results of MTB/RIF or Xpert Ultra can be used as a decision-making tool in the following ways:
M tuberculosis detected/rifampicin resistance not detected: child would start treatment for drug-sensitive tuberculosis;
M tuberculosis detected/rifampicin resistance detected: child would need further resistance testing and would start treatment for drug-resistant tuberculosis according to the country guidelines;
M tuberculosis not detected: a negative Xpert MTB/RIF or Xpert Ultra result does not rule out tuberculosis disease. Therefore, clinicians should still consider initiation of tuberculosis treatment in children with history and clinical features suggestive of tuberculosis disease despite a negative Xpert MTB/RIF or Xpert Ultra result. A negative Xpert MTB/RIF or Xpert Ultra result may also represent a true negative.
Possible consequences of a false-positive and a false-negative result may include the following:
false positives (FP): children and their families would likely experience anxiety and morbidity caused by additional testing, unnecessary treatment, and possible adverse effects; possible stigma associated with a tuberculosis or drug-resistant tuberculosis diagnosis and the chance that a false positive may halt further diagnostic evaluation;
false negatives (FN): would likely result in an increased risk of morbidity and mortality and delayed treatment initiation for patients.

METHODS

Search methods

We searched the following databases: Cochrane Infectious Diseases Group Specialized Register; MEDLINE (OVID, from 1966); Embase (OVID, from 1974); Science Citation Index - Expanded (from 1900), Conference Proceedings Citation Index - Science (CPCI-S, from 1990), and BIOSIS Previews (from 1926), all three from the Web of Science; Scopus (Elsevier, from 1970); and Latin American Caribbean Health Sciences Literature (LILACS) (BIREME, from 1982). We also searched ClinicalTrials.gov, the WHO International Clinical Trials Registry (ICTRP) Platform (www.who.int/trialsearch), and the International Standard Randomized Controlled Trials Number (ISRCTN) registry (www.isrctn.com/) for trials in progress, and ProQuest Dissertations & Theses A&I (from 1990) for dissertations to 29 April 2019, without language restriction.

Selection criteria

We included cross-sectional and cohort studies that evaluated the accuracy of Xpert MTB/RIF and Xpert Ultra for the diagnosis of active tuberculosis and rifampicin resistance in children 0 to 14 years of age with signs and symptom of tuberculosis. For tuberculosis detection, we used a microbiological reference standard (culture) and a composite reference standard. We defined the composite reference standard as a positive culture or a clinical decision to initiate treatment for

tuberculosis. In the absence of information on tuberculosis treatment, for the composite reference standard, we accepted a study specific definition (i.e. a definition of confirmed tuberculosis defined by the primary study authors), if available. For rifampicin resistance detection, the reference standards were culture-based drug susceptibility testing and MTBDR*plus*. Studies were eligible for inclusion if they described the use of Xpert MTB/RIF or Xpert Ultra on routine respiratory specimens, such as expectorated or induced sputum, gastric specimens, and nasopharyngeal specimens. In addition, we included studies evaluating stool because tuberculosis bacilli are present in swallowed sputum and recoverable from stool specimens using Xpert assays. For extrapulmonary TB, we included studies that assessed the accuracy of the index tests for TB meningitis and lymph node TB against a microbiological or composite reference standard. References to included studies are included in Appendix A.

Data collection and analysis
Two review authors independently extracted data using a standardized form. We assessed study quality using QUADAS-2 and performed meta-analyses using a bivariate random-effects model to determine summary estimates of sensitivity and specificity for Xpert MTB/RIF and Xpert Ultra separately for the different specimens used for detection of tuberculosis and rifampicin resistance. We investigated potential sources of heterogeneity by reference standard and by clinical subgroup, including smear result and HIV infection status.

Assessment of certainty of the evidence

We assessed the certainty of the evidence using the GRADE approach, and GRADEpro Guideline Development Tool (GDT) software (GRADEpro GDT 2015). For each outcome, we considered the certainty of the evidence to begin as high when high-quality observational studies (cross-sectional or cohort studies) enrolled participants with diagnostic uncertainty. If we had a reason for downgrading, we used our judgement to classify the reason as serious (downgraded by one level) or very serious (downgraded by two levels).

We applied GRADE in the following ways:

Risk of bias: we used QUADAS-2 to assess risk of bias.

Indirectness: we used QUADAS-2 for concerns of applicability and looked for important differences between the populations studied (for example, the spectrum of disease), the setting, index test, and outcomes, and asked whether differences were sufficient to lower certainty in results.

Inconsistency: GRADE recommends downgrading for unexplained inconsistency in sensitivity and specificity estimates. We carried out prespecified analyses and did not downgrade when we believed we could explain inconsistency in the accuracy estimates.

Imprecision: we considered a precise estimate to be one that would allow a clinically meaningful decision. We considered the width of the CI and asked ourselves, 'Would we make a different decision if the lower or upper boundary of the CI represented the truth?'. In addition, we calculated projected ranges for true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP) for a given prevalence of tuberculosis and made judgements on imprecision from these calculations.

Publication bias: we rated publication bias as undetected (not serious) because of the comprehensiveness of the literature search and following extensive outreach to tuberculosis

researchers to identify studies. As we included a large number of studies, we thought that had we missed several small studies, the results would probably not be different.
RESULTS

Pico 2: Detection of Pulmonary TB in Children
The initial search resulted in 835 individual records, with one additional reference identified through other sources, from which 701 were excluded. We retrieved 134 articles, and after full-text review, included 50 studies in the quantitative meta-analysis. The PRISMA diagram and reasons for exclusion are included in Appendix C.

Of the 50 studies included in the review, 40 (80%) took place in high tuberculosis burden and 40 in high TB/HIV burden countries. For pulmonary tuberculosis detection, we included 43 studies that evaluated the diagnostic accuracy of Xpert MTB/RIF in children and three studies that evaluated both Xpert Ultra and Xpert MTB/RIF. 42 studies evaluated pulmonary tuberculosis using a reference standard of culture and one study evaluated pulmonary tuberculosis using smear only. The results of this large, single study evaluating pulmonary tuberculosis using a reference standard of smear are described in Appendix D.

Concerning the assessment of methodological quality, in the patient selection domain, we judged most studies (81%) evaluating pulmonary tuberculosis to have low risk of bias if they consecutively recruited patients. In the index test domain, we judged all studies to have low risk of bias based on the automated nature of the Xpert MTB/RIF test. In the flow and timing domain we judged most studies (88%) to have low risk of bias according to the timing of the index test and reference test specimen collection. In the reference standard domain, with respect to the microbiological reference standard, we judged 46% of studies to have unclear risk of bias because only one culture was used to exclude tuberculosis. With respect to the composite reference standard, we judged all studies to have unclear risk of bias because of imperfect accuracy of the composite reference standard and differing definitions of this standard used by the primary study authors. Regarding applicability, in the patient selection domain, we judged high or unclear concern for 49% of studies because participants were evaluated exclusively as inpatients at tertiary care centres, or we were not sure about the clinical setting. With respect to applicability of the index test, we judged most studies (72%) as having low concern owing to standardized application of the index tests. We judged 11 studies evaluating stool as a specimen for Xpert MTB/RIF or Xpert Ultra to have unclear risk of bias because of the absence of a standardized protocol for stool preparation. With respect to applicability of the reference standard, we considered this to be a low concern for most studies (91%).

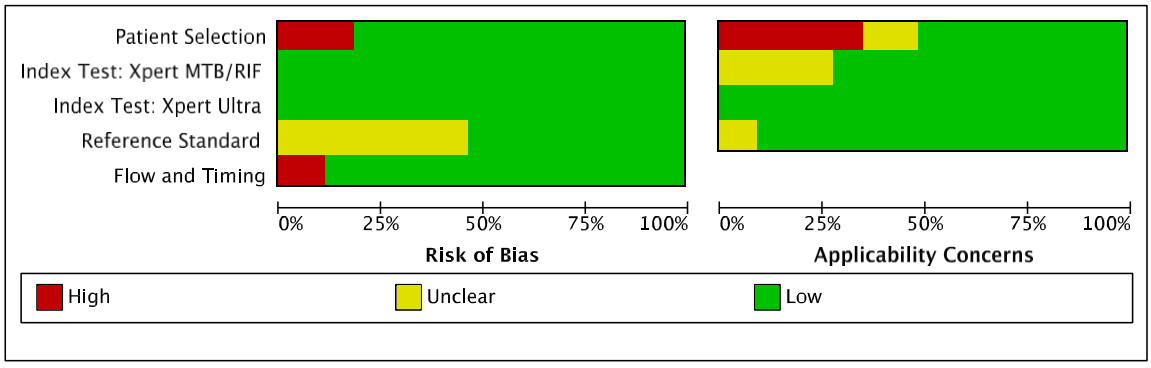


Figure 1. Risk of bias and applicability concerns graph for pulmonary TB studies: review authors' judgements about each domain presented as percentages across included studies.

For the meta-analysis, a total of 23 studies (6703 participants) evaluated sputum specimens; 14 studies (3482 participants) evaluated gastric specimens; four studies (1125 participants) evaluated nasopharyngeal specimens; 11 studies (1592 participants) evaluated stool specimens; all of the above evaluated Xpert MTB/RIF alone. No studies evaluated Xpert Ultra alone. Three studies (753 participants) evaluated both Xpert MTB/RIF and Xpert Ultra on frozen sputum specimens. One study (195 participants) evaluated both Xpert MTB/RIF and Xpert Ultra on nasopharyngeal specimens.

For the WHO Guideline meeting, our review team addressed PICO 2 and PICO 4 and several subquestions of PICO 5. The specific findings follow below.

PICO 2 Subquestions

16. Should Xpert MTB/RIF be used to diagnose pulmonary TB in sputum in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 64.6% (55.3 to 72.9) and 99.0% (98.1 to 99.5), (23 studies, 6703 participants; moderate-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 74 would be Xpert MTB/RIF-positive and 9 (12%) would not have tuberculosis (false-positives); 926 would be Xpert MTB/RIF-negative and 35 (4%) would have tuberculosis (false-negatives).

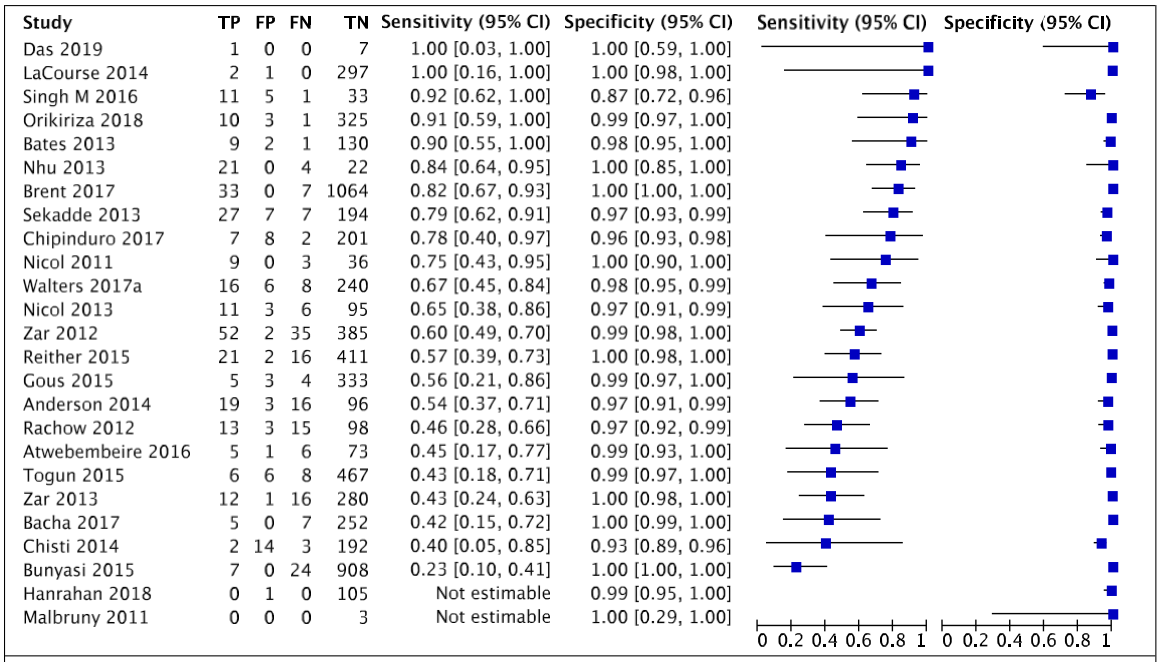


Figure 2. Q16 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum against a microbiologic reference standard and stratified by the number of cultures obtained. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

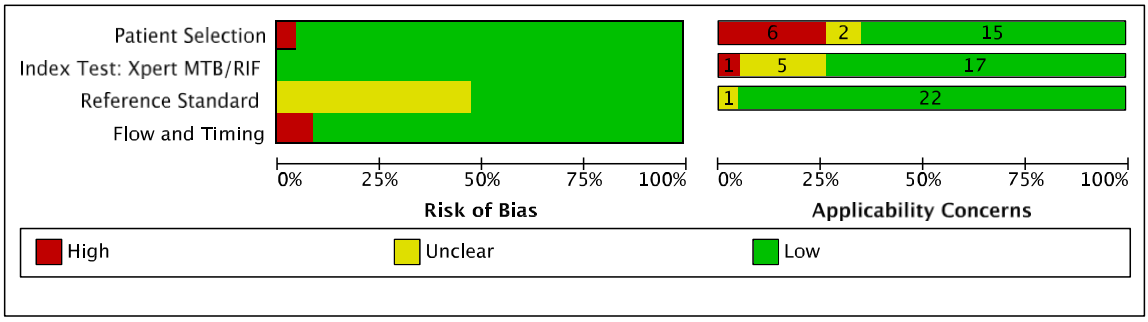


Figure 3. Risk of bias and applicability concerns graph for pulmonary TB studies using sputum: review authors' judgements about each domain presented as percentages across included studies.

17. Should Xpert MTB/RIF be used to diagnose pulmonary TB in sputum in children with signs and symptoms of pulmonary TB, against a composite reference standard?
Xpert MTB/RIF pooled sensitivity and specificity were 19.7% (12.1 to 30.4) and 100% (100 to 100), (16 studies, 4379 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 smear-negative children where 100 have pulmonary tuberculosis by a microbiologic reference standard, 20 would be Xpert MTB/RIF-positive and 0 (0%) would not have tuberculosis (false-positives); 980 would be Xpert MTB/RIF-negative and 80 (8%) would have tuberculosis (false-negatives).

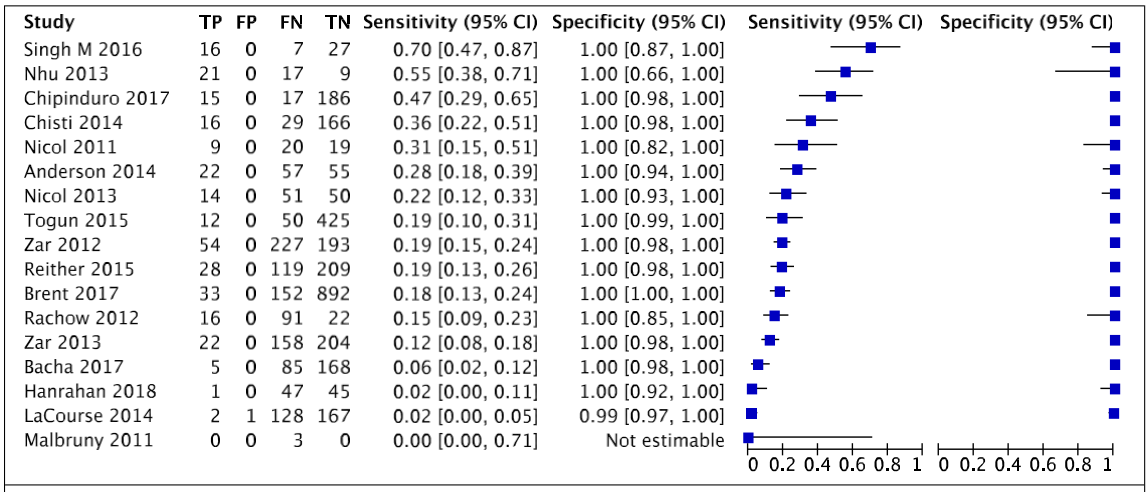


Figure 4. Q17 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum against a composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

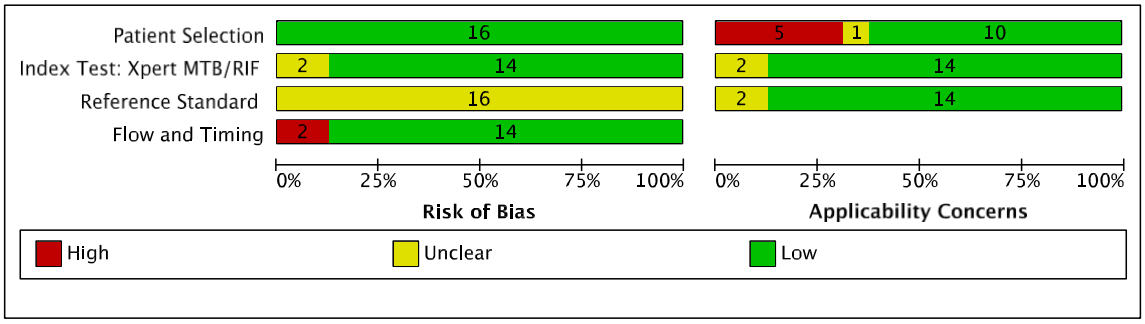


Figure 5. Risk of bias and applicability concerns graph for pulmonary TB studies using sputum compared to a composite reference standard: review authors' judgements about each domain presented as percentages across included studies.

18. Should Xpert Ultra be used to diagnose pulmonary TB in sputum in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?
Xpert Ultra pooled sensitivity and specificity were 72.8% (64.7 to 79.6) and 97.5% (95.8 to 98.5), (3 studies, 697 participants; low-certainty evidence for sensitivity; high-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 100 would be Xpert Ultra-positive and 27 (27%) would not have tuberculosis (false-positives); 900 would be Xpert Ultra-negative and 27 (3%) would have tuberculosis (false-negatives).

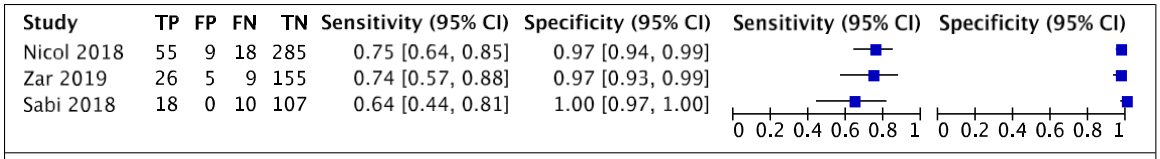


Figure 6. Q18 Forest plots of Xpert Ultra sensitivity and specificity for pulmonary tuberculosis using sputum against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

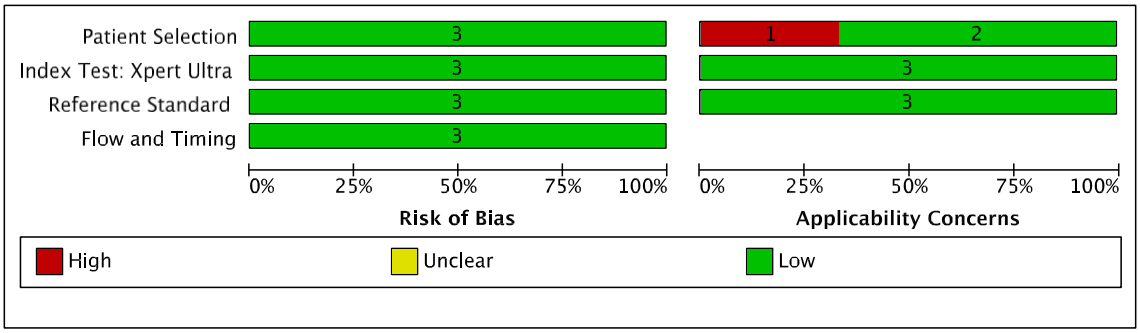


Figure 7. Risk of bias and applicability concerns graph for pulmonary TB studies on Xpert Ultra using sputum: review authors' judgements about each domain presented as percentages across included studies.

19. Should Xpert Ultra be used to diagnose pulmonary TB in sputum in children with signs and symptoms of pulmonary TB, against a composite reference standard?

Xpert Ultra pooled sensitivity and specificity were 23.5% (20.0 to 27.4) and 99.2% (96.9 to 99.8), (3 studies, 753 participants; low-certainty evidence for sensitivity; low-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis by a composite reference standard, 33 would be Xpert Ultra-positive and 9 (27%) would not have tuberculosis (false-positives); 967 would be Xpert Ultra-negative and 76 (8%) would have tuberculosis (false-negatives).

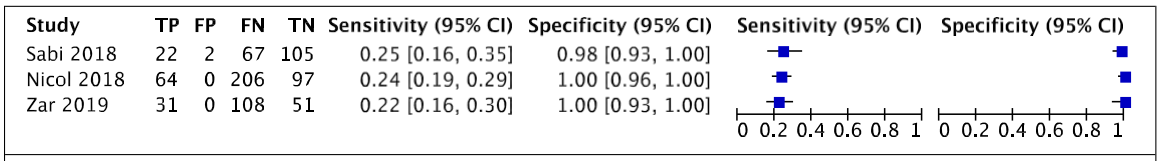


Figure 8. Q19 Forest plots of Xpert Ultra sensitivity and specificity for pulmonary tuberculosis using sputum against a composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

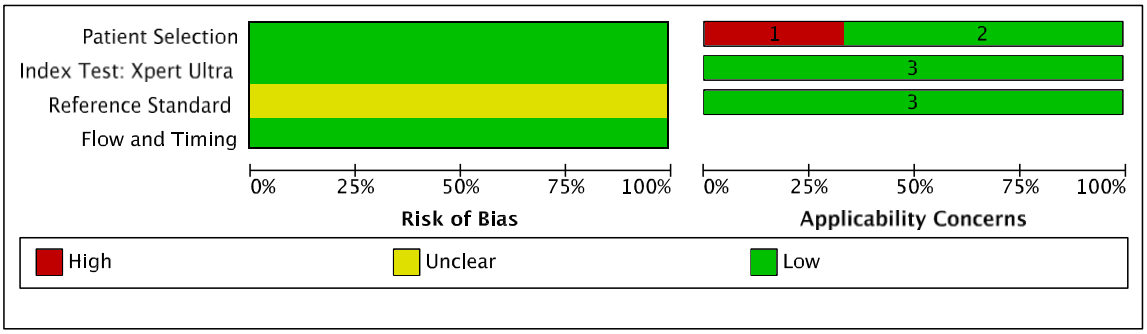


Figure 9. Risk of bias and applicability concerns graph for pulmonary TB studies on Xpert Ultra against a composite reference standard using sputum: review authors' judgements about each domain presented as percentages across included studies.

20. Should Xpert MTB/RIF be used to diagnose pulmonary TB in sputum in smear-positive children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

There were only 91 cases among smear-positive children and 103 total patients (11 studies, 103 participants). The bivariate model did not converge for this analysis. We performed a univariate analysis for sensitivity, which was 97.8% (91.6 to 99.4) in smear-positive children.

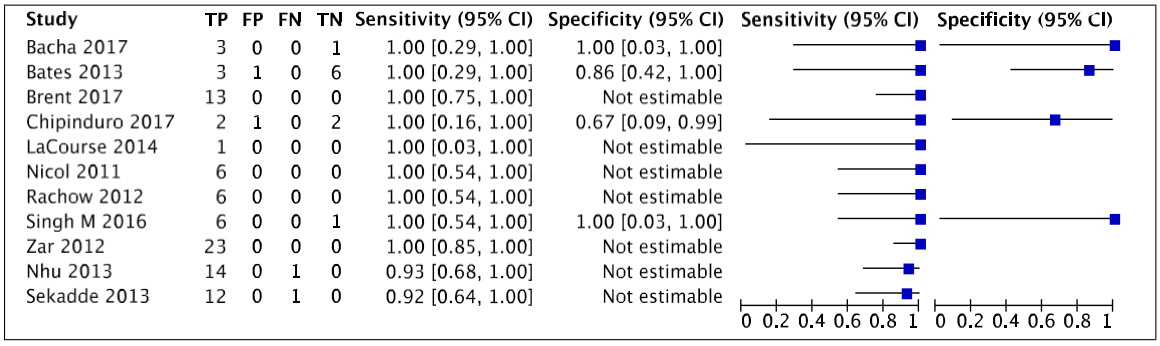


Figure 10. Q20 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum against a microbiologic reference standard in smear positive patients. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

21. Should Xpert MTB/RIF be used to diagnose pulmonary TB in sputum in smear-negative, culture-positive children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 58.9% (45.6 to 71.0) and 99.1% (97.1 to 99.7), (12 studies, 3118 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 children where 100 have smear-negative pulmonary tuberculosis on culture, 68 would be Xpert MTB/RIF-positive and 9 (13%) would not have tuberculosis (false-positives); 932 would be Xpert MTB/RIF-negative and 41 (4%) would have tuberculosis (false-negatives).

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Malbruny 2011	0	0	0	3	Not estimable	1.00 [0.29, 1.00]
LaCourse 2014	1	1	0	297	1.00 [0.03, 1.00]	1.00 [0.98, 1.00]
Bates 2013	6	1	1	124	0.86 [0.42, 1.00]	0.99 [0.96, 1.00]
Singh M 2016	5	5	1	32	0.83 [0.36, 1.00]	0.86 [0.71, 0.95]
Brent 2017	20	0	7	1064	0.74 [0.54, 0.89]	1.00 [1.00, 1.00]
Chipinduro 2017	5	7	2	199	0.71 [0.29, 0.96]	0.97 [0.93, 0.99]
Nhu 2013	7	0	3	22	0.70 [0.35, 0.93]	1.00 [0.85, 1.00]
Sekadde 2013	14	7	6	194	0.70 [0.46, 0.88]	0.97 [0.93, 0.99]
Nicol 2011	3	0	3	36	0.50 [0.12, 0.88]	1.00 [0.90, 1.00]
Zar 2012	29	2	35	385	0.45 [0.33, 0.58]	0.99 [0.98, 1.00]
Chisti 2014	2	14	3	192	0.40 [0.05, 0.85]	0.93 [0.89, 0.96]
Rachow 2012	7	3	15	98	0.32 [0.14, 0.55]	0.97 [0.92, 0.99]
Bacha 2017	2	0	7	251	0.22 [0.03, 0.60]	1.00 [0.99, 1.00]

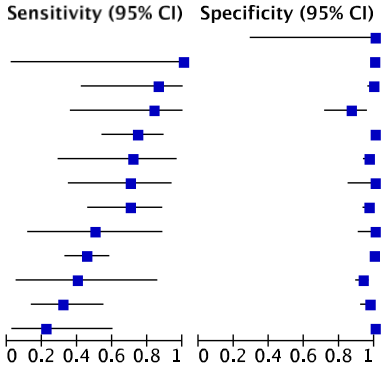


Figure 11. Q21 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum against a microbiologic reference standard in smear negative patients. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

22. Should Xpert MTB/RIF be used to diagnose pulmonary TB in sputum in HIV-positive children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 72.2% (59.9 to 81.8) and 99.4% (97.2 to 99.9) (10 studies, 642 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 HIV-positive children where 100 have pulmonary tuberculosis on culture, 81 would be Xpert MTB/RIF-positive and 9 (11%) would not have tuberculosis (false-positives); 919 would be Xpert MTB/RIF-negative and 28 (3%) would have tuberculosis (false-negatives).

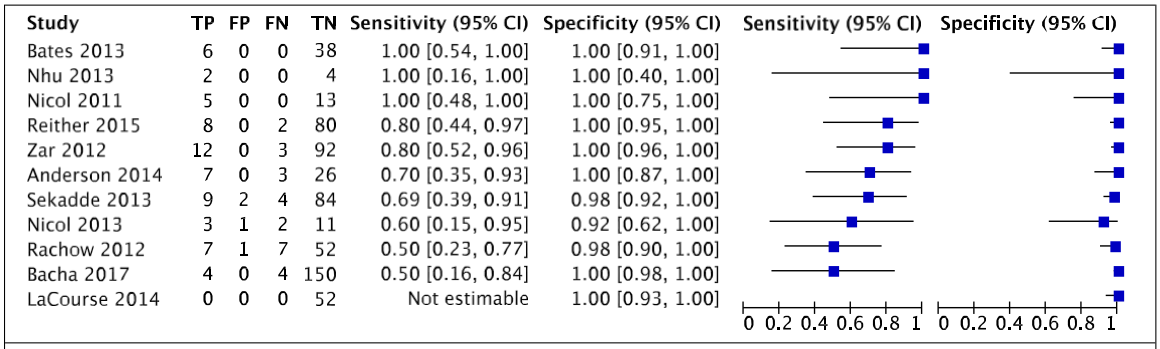


Figure 12. Q22 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum against a microbiologic reference standard in HIV positive patients. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

23. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a gastric specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 73.0% (52.9 to 86.7) and 98.1% (95.5 to 99.2), (14 studies, 3482 participants; very low-certainty evidence for sensitivity; low certainty of evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 90 would be Xpert MTB/RIF-positive and 17 (19%) would not have tuberculosis (false-positives); 910 would be Xpert MTB/RIF-negative and 27 (3%) would have tuberculosis (false-negatives).

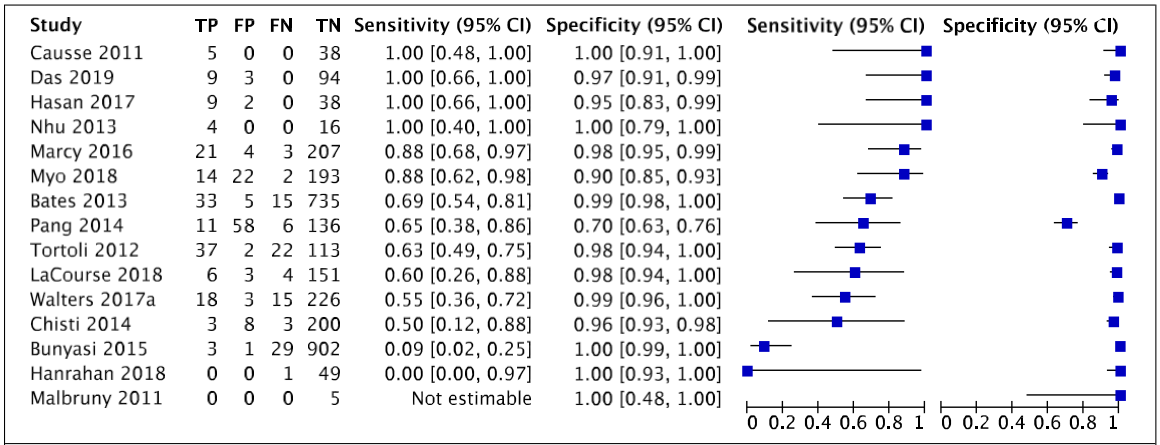


Figure 13. Q23 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a gastric specimen against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

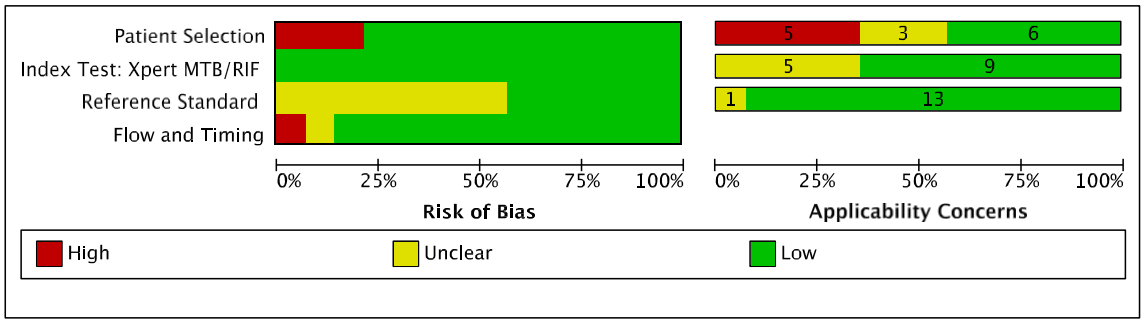


Figure 14. Risk of bias and applicability concerns graph for pulmonary TB studies using gastric specimens: review authors' judgements about each domain presented as percentages across included studies.

24. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a gastric specimen in children with signs and symptoms of pulmonary TB, against a composite reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 31.7% (20.2 to 46.0) and 99.7% (97.1 to 100), (6 studies, 933 participants; very low-certainty evidence for sensitivity; moderate certainty for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis by the composite reference standard,41 would be Xpert MTB/RIF-positive and 9 (22%) would not have tuberculosis (false-positives); 959 would be Xpert MTB/RIF-negative and 68 (7%) would have tuberculosis (false-negatives).

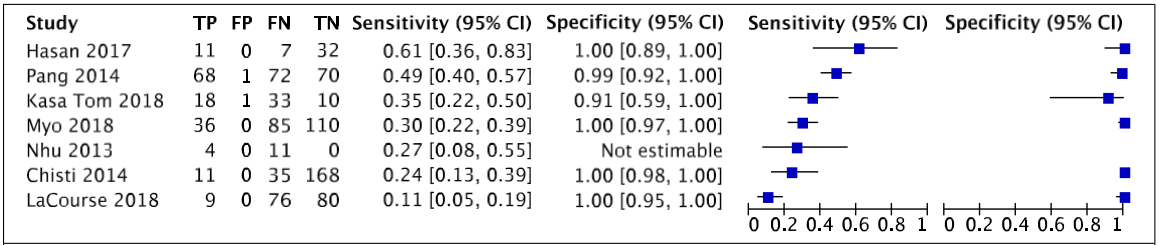


Figure 15. Q24 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a gastric specimen against a composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

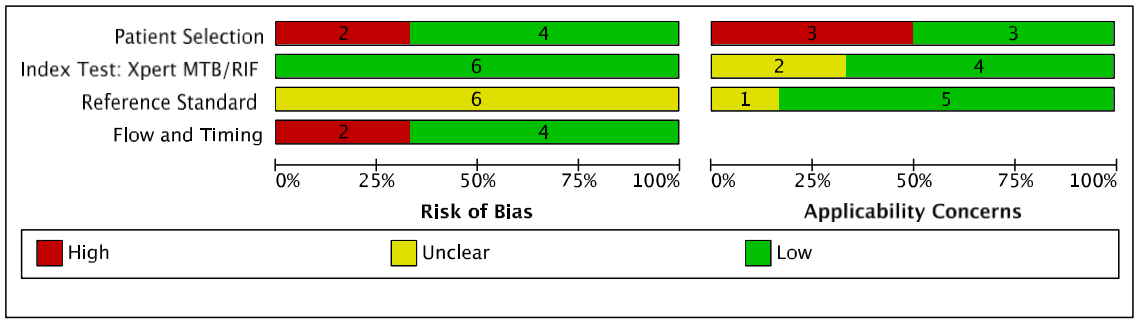


Figure 16. Risk of bias and applicability concerns graph for pulmonary TB studies using gastric specimens against a composite reference standard: review authors' judgements about each domain presented as percentages across included studies.

25. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a gastric specimen in HIV-positive children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 73.3% (54.9 to 86.1) and 98.5% (97.1 to 99.2) (3 studies, 634 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 82 would be Xpert MTB/RIF-positive and 9 (11%) would not have tuberculosis (false-positives); 918 would be Xpert MTB/RIF-negative and 27 (3%) would have tuberculosis (false-negative).

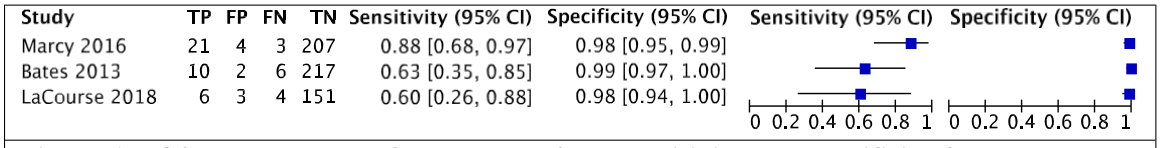


Figure 17. Q25 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a gastric specimen against a microbiologic reference standard in HIV positive children. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

26. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a nasopharyngeal specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 45.7% (27.6 to 65.1) and 99.6% (98.9 to 99.8), (4 studies, 1125 participants; moderate-certainty evidence for sensitivity; high-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 46 would be Xpert MTB/RIF-positive and 0 (0%) would not have tuberculosis (false-positives); 954 would be Xpert MTB/RIF-negative and 54 (6%) would have tuberculosis (false-negatives).

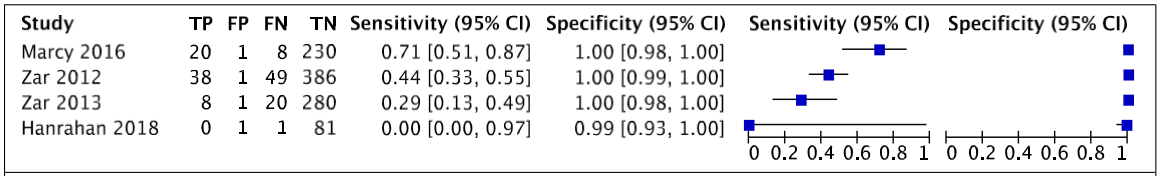


Figure 18. Q26 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a nasopharyngeal specimen against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

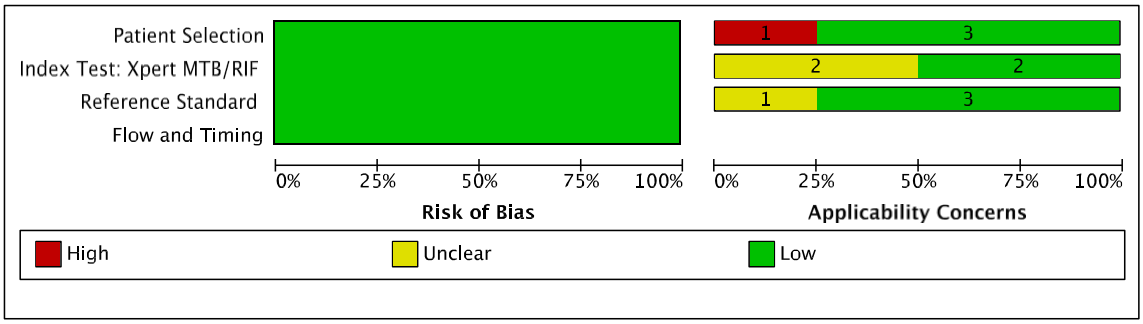


Figure 19. Risk of bias and applicability concerns graph for pulmonary TB studies using nasopharyngeal specimens: review authors' judgements about each domain presented as percentages across included studies.

27. Should Xpert Ultra be used to diagnose pulmonary TB in a nasopharyngeal specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF sensitivity and specificity were 45.7% (28.9 to 63.3)) and 97.5% (93.7 to 99.3), (1 study, 195 participants; very low-certainty evidence for sensitivity; low-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 46 would be Xpert MTB/RIF-positive and 0 (0%) would not have tuberculosis (false-positives); 954 would be Xpert MTB/RIF-negative and 54 (6%) would have tuberculosis (false-negatives).

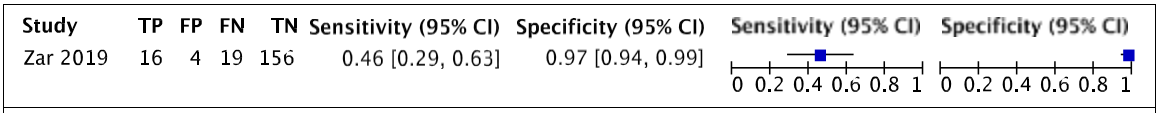


Figure 20. Q27 Forest plots of Xpert Ultra sensitivity and specificity for pulmonary tuberculosis using a nasopharyngeal specimen against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

28. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a stool specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 61.5% (44.1 to 76.4) and 98.5% (97.0 to 99.2), (11 studies, 1592 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis on culture, 75 would be Xpert MTB/RIF-positive and 13 (17%) would not have tuberculosis (false-positives); 925 would be Xpert MTB/RIF-negative and 38 (4%) would have tuberculosis (false-negatives).

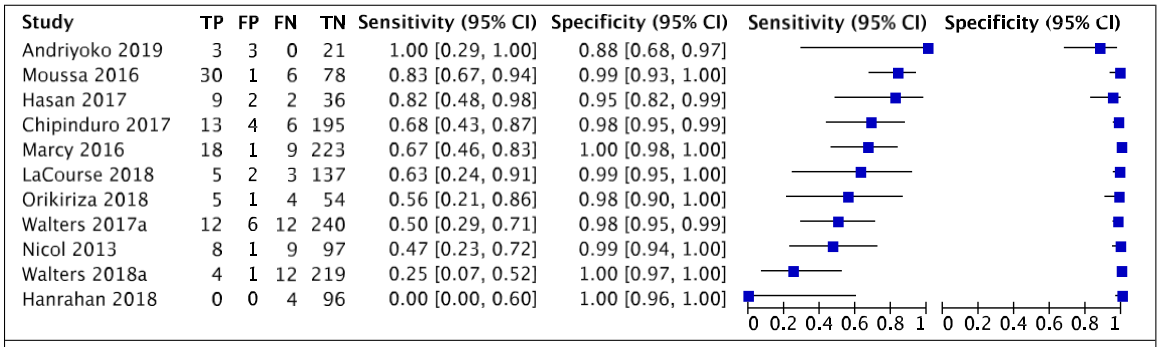


Figure 21. Q28 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a stool specimen against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

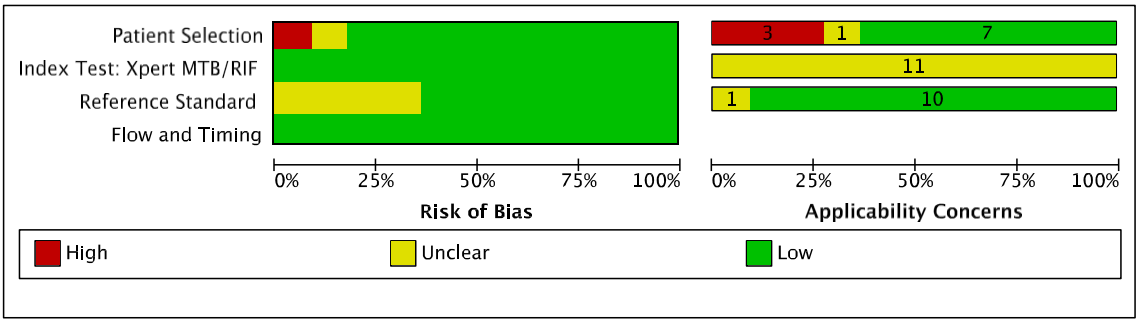


Figure 22. Risk of bias and applicability concerns graph for pulmonary TB studies using stool specimens: review authors' judgements about each domain presented as percentages across included studies.

29. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a stool specimen in children with signs and symptoms of pulmonary TB, against a composite reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 16.3% (8.4 to 29.2) and 99.7% (97.8 to 100), (10 studies, 1739 participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 children where 100 have pulmonary tuberculosis by a composite reference standard, 25 would be Xpert MTB/RIF-positive and 9 (36%) would not have tuberculosis (false-positives); 975 would be Xpert MTB/RIF-negative and 84 (9%) would have tuberculosis (false-negatives).

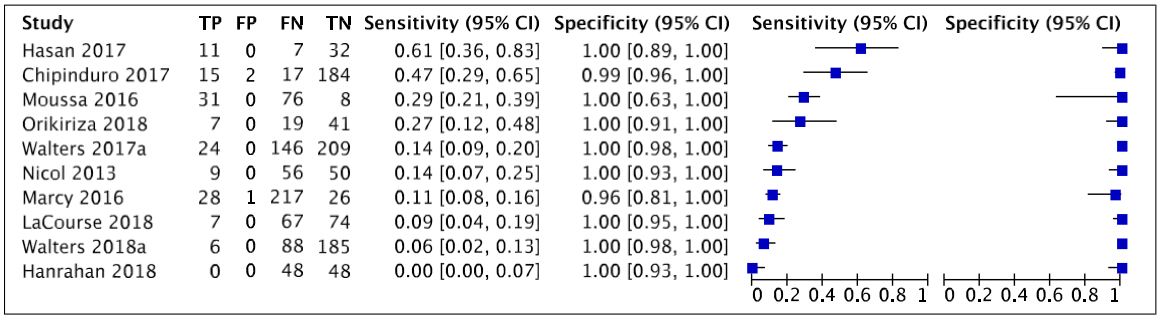


Figure 23. Q29 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a stool specimen against a composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

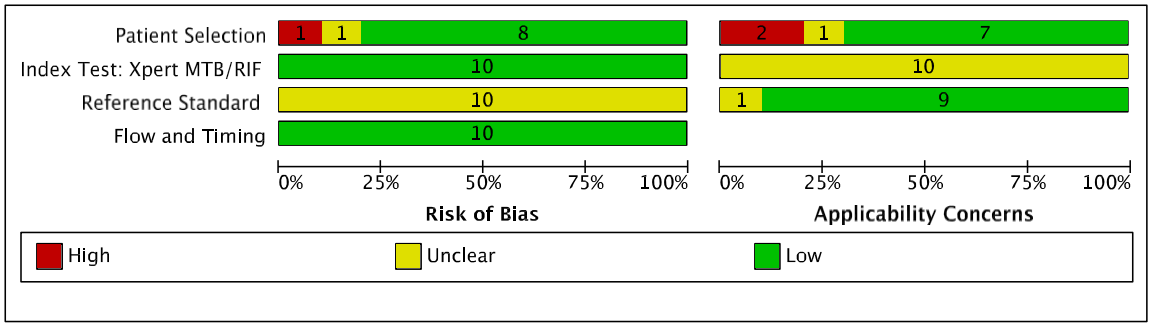


Figure 24. Risk of bias and applicability concerns graph for pulmonary TB studies using stool specimens against a composite reference standard: review authors' judgements about each domain presented as percentages across included studies.

30. Should Xpert MTB/RIF be used to diagnose pulmonary TB in a stool specimen in HIV-positive children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 69.8% (56.3 to 80.6) and 98.6% (96.1 to 99.5), (4 studies, 526 participants; low-certainty evidence for sensitivity; high-certainty evidence for specificity).

For a population of 1000 HIV-positive children where 100 have pulmonary tuberculosis on culture, 88 would be Xpert MTB/RIF-positive and 18 (20%) would not have tuberculosis (false-positives); 912 would be Xpert MTB/RIF-negative and 30 (3%) would have tuberculosis (false-negatives).

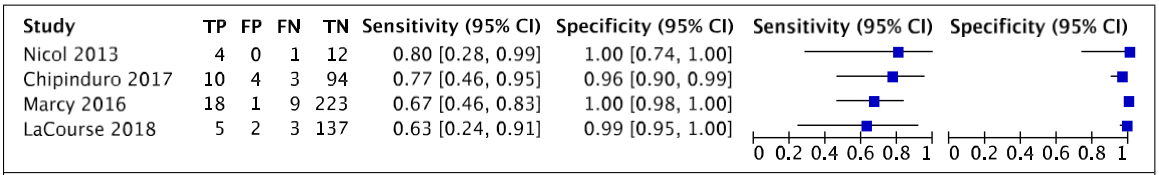


Figure 25. Q30 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using a stool specimen against a microbiologic reference standard in HIV-positive patients. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

Table 1. PICO 2 summary: diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for pulmonary TB and rifampicin resistance in children*

PICO, sub-question	Test, analysis group	Reference Standard	Studies	Number of children (TB cases)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive Predictive Value % (95% CI)	Negative Predictive Value % (95% CI)
16	Xpert MTB/RIF, sputum	MRS	23	6703 (494)	64.6 (55.3 to 72.9)	99.0 (98.1 to 99.5)	88.2 (79.6 to 93.5)	96.2 (95.1 to 97.0)
17	Xpert MTB/RIF, sputum	CRS	16	4379 (1541)	19.7 (12.1 to 30.4)	100 (99.8 to 100)	98.4 (89.2 to 99.8)	91.8 (90.9 to 92.6)
18	Xpert Ultra, sputum	MRS	3	697 (136)	72.8 (64.7 to 79.6)	97.5 (95.8 to 98.5)	76.4 (65.6 to 84.6)	97.7 (95.9 to 97.7)
19	Xpert Ultra, sputum	CRS	3	753 (498)	23.5 (20.0 to 27.4)	99.2 (96.9 to 99.8)	76.9 (45.3 to 93.0)	92.1 (91.7 to 92.5)
20**	Xpert MTB/RIF, sputum, smear positive	MRS	11	91 (88)	97.8 (91.6 to 99.4)	–	–	–
21	Xpert MTB/RIF, sputum, smear negative	MRS	12	3118 (184)	58.9 (45.6 to 71.0)	99.1 (97.1 to 99.7)	88.4 (68.8 to 96.3)	95.6 (94.0 to 96.8)
22	Xpert MTB/RIF, sputum, HIV positive	MRS	10	642 (88)	72.2 (59.9 to 81.8)	99.4 (97.2 to 99.9)	93.2 (74.0 to 98.5)	97.0 (95.5 to 97.9)
23	Xpert MTB/RIF, gastric specimen	MRS	14	3482 (273)	73.0 (52.9 to 86.7)	98.1 (95.5 to 99.2)	81.0 (65.5 to 90.6)	97.0 (94.5 to 98.4)
24	Xpert MTB/RIF, gastric specimen	CRS	6	933 (461)	31.7 (20.2 to 46.0)	99.7 (97.1 to 100)	91.7 (58.3 to 98.9)	92.9 (91.6 to 94.0)
25	Xpert MTB/RIF, gastric specimen, HIV positive	MRS	3	634 (50)	73.3 (54.9 to 86.1)	98.5 (97.1 to 99.2)	84.1 (72.7 to 91.3)	97.1 (93.8 to 98.4)
26	Xpert MTB/RIF, nasopharyngeal specimen	MRS	4	1125 (144)	45.7 (27.6 to 65.1)	99.6 (98.9 to 99.8)	92.6 (81.1 to 97.3)	94.3 (92.0 to 95.9)
27	Xpert Ultra, nasopharyngeal specimen	MRS	1	195 (35)	45.7 (28.9 to 63.3)	97.5 (93.7 to 99.3)	67.0 (42.0 to 85.1)	94.1 (92.2 to 95.6)
28	Xpert MTB/RIF, stool specimen	MRS	11	1592 (174)	61.5 (44.1 to 76.4)	98.5 (97.0 to 99.2)	81.7 (72.2 to 88.5)	95.8 (93.8 to 97.3)

PICO, sub-question	Test, analysis group	Reference Standard	Studies	Number of children (TB cases)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive Predictive Value % (95% CI)	Negative Predictive Value % (95% CI)
29	Xpert MTB/RIF, stool specimen	CRS	10	1739 (879)	16.3 (8.43 to 29.2)	99.7 (97.8 to 100)	87.4 (42.8 to 98.5)	91.5 (90.5 to 92.4)
30	Xpert MTB/RIF, stool specimen, HIV positive	MRS	4	526 (53)	69.8 (56.3 to 80.6)	98.6 (96.1 to 99.5)	84.7 (66.2 to 94.0)	96.7 (95.1 to 97.8)
31	Xpert MTB/RIF, rifampicin resistance	Culture-DST, MTBDR*plus*	6	223 (20)	90.0 (67.6 to 97.5)	98.3 (87.7 to 99.8)	85.7 (42.7 to 98.0)	98.9 (95.9 to 99.7)

* Predictive values were determined at a pre-test probability of 10%

Abbreviations: CI: confidence interval; CRS: composite reference standard (culture, tuberculosis treatment initiation and clinically diagnosed tuberculosis); DST: drug susceptibility testing; MRS: microbiological reference standard (culture)

**We performed a univariate meta-analysis for this analysis group.

PICO 2: Detection of Rifampicin Resistance in Children

We identified 14 studies that provided data on the detection of rifampicin resistance. Nevertheless, we were able to include only six studies in the meta-analysis to generate evidence about the detection of rifampicin resistance. All of the six studies (223 participants) evaluated only Xpert MTB/RIF, and were conducted in high tuberculosis burden countries and in high MDR TB burden countries. 42% of studies had a low risk of bias with respect to patient selection while all studies had a low risk of bias with respect to the reference standard. Risk of bias was considered low for the reference standard (86%) if an automated process was used or it was clear that the reference standard results were interpreted without knowledge of the index tests. Nine studies (64%) had high or uncertain applicability concerns regarding patient selection due to enrolment exclusively from inpatient or tertiary centers.

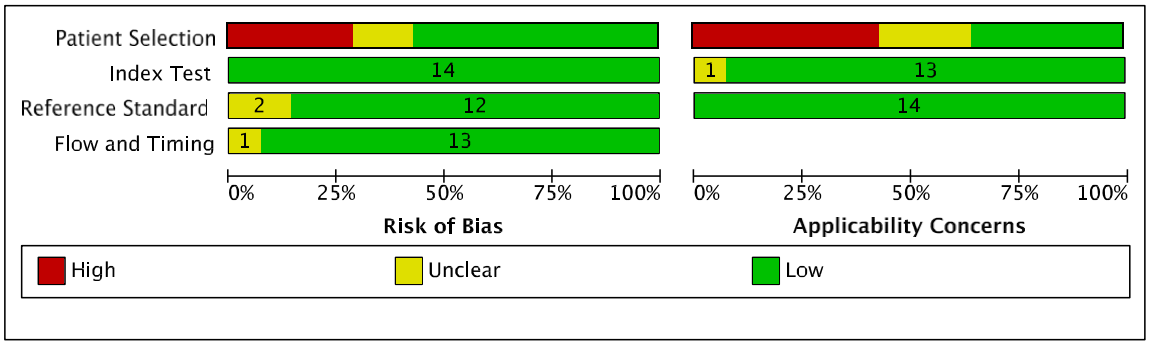


Figure 26. Risk of bias and applicability concerns graph for rifampicin resistance: review authors' judgements about each domain presented as percentages across included studies.

31. Should Xpert MTB/RIF be used to diagnose rifampicin resistance in children with signs and symptoms of pulmonary TB?

Xpert MTB/RIF pooled sensitivity (95% CI) and specificity (95% CI) were 90.0% (67.6% to 97.5%) and 98.3% (87.7% to 99.8%), (6 studies, 223 unselected participants; low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

For a population of 1000 children where 100 have rifampicin resistance, 108 would be Xpert MTB/RIF-rifampicin resistance detected and 18 (17%) would not have rifampicin resistance (false-positives); 892 would be Xpert MTB/RIF-rifampicin resistance NOT detected and 10 (1%) would have rifampicin resistance (false-negatives).

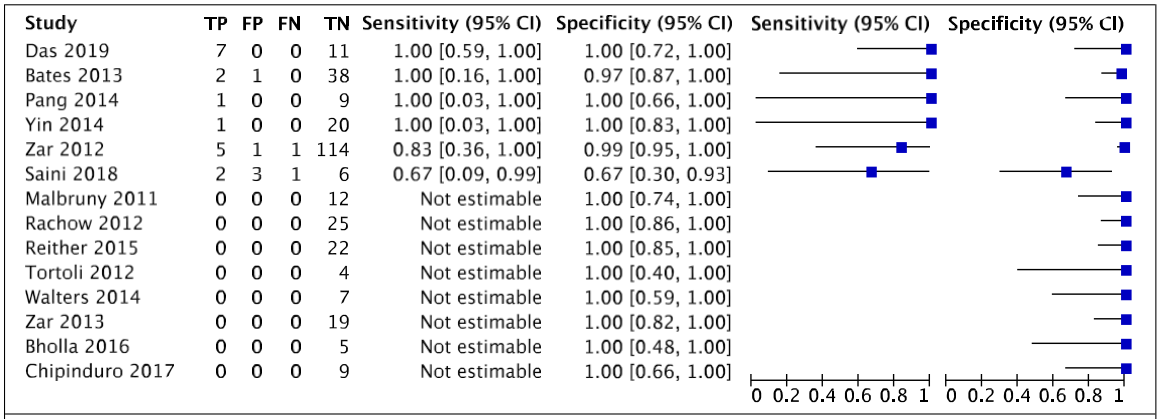


Figure 27. Q31 Forest plots of Xpert MTB/RIF sensitivity and specificity for rifampicin resistance against a reference standard of phenotypic drug susceptibilities or line probe

resistance assays. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

PICO 4: Detection of Extrapulmonary Tuberculosis in Children

To evaluate detection of extrapulmonary tuberculosis, we included studies that evaluated the diagnostic accuracy of Xpert MTB/RIF in children with signs or symptoms of lymph node TB or TB meningitis. We did not identify any studies that evaluated the accuracy of Xpert Ultra for detecting lymph node TB or TB meningitis.

For diagnosis of lymph node TB, we included six studies (210 participants) evaluating Xpert MTB/RIF against a microbiologic reference standard of smear or culture on lymph node specimens. Two studies (105 participants) evaluated Xpert MTB/RIF against a composite reference standard for lymph node TB. For TB meningitis, we included six studies (241 participants) evaluating Xpert MTB/RIF against culture on cerebrospinal fluid. In addition, two studies (155 participants) assessed Xpert MTB/RIF against a composite reference standard including a clinical diagnosis of TB meningitis.

Of the 13 studies evaluating extrapulmonary TB, in the patient selection domain, we considered nine (69%) to have low risk of bias because of prospective consecutive enrolment. In the reference standard domain, with respect to the microbiological reference standard, we judged 12 studies (92%) to have unclear risk of bias because only one culture was obtained. For the composite reference standard, we judged all studies to have unclear risk of bias because of imperfect accuracy of the composite reference standard and differing definitions of this standard used by the primary study authors. In the flow and timing domain, we judged one study (11%) to have unclear risk of bias. For applicability, in the patient selection domain, we judged three studies (33%) to have unclear concern and five studies (38%) to have high concern. In the index test domain, we judged all studies to have low concern. And in the reference standard domain, we judged one study (8%) to have unclear concern and one study to have high concern (8%) and the remaining studies to have low concern.

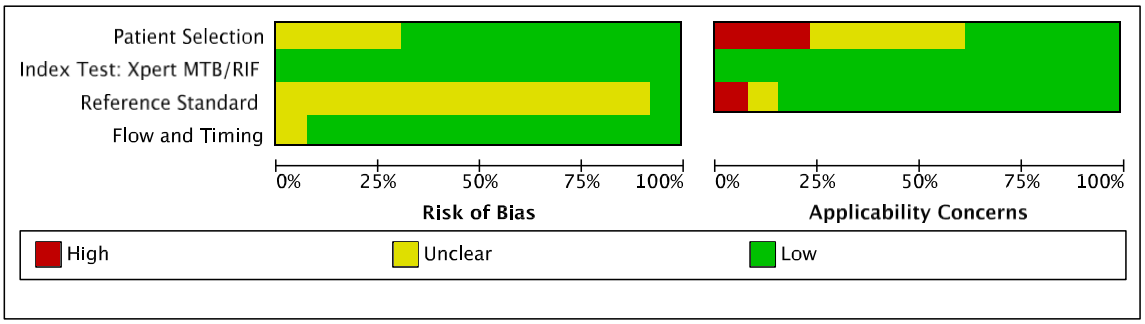


Figure 28. Risk of bias and applicability concerns graph for extrapulmonary TB: review authors' judgements about each domain presented as percentages across included studies.

PICO 4 Subquestions

57. Should Xpert MTB/RIF be used to diagnose TB meningitis in CSF in children with signs and symptoms of TB meningitis, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 54.0% (27.8 to 78.2) and 93.8% (84.5 to 97.6) (6 studies, 262 participants; very low-certainty evidence for sensitivity; low certainty evidence for specificity).

For a population of 1000 children where 100 have TB meningitis on culture, 86 would be Xpert MTB/RIF-positive and 59 (69%) would not have tuberculosis (false-positives); 914 would be Xpert MTB/RIF-negative and 23 (3%) would have tuberculosis (false-negatives).

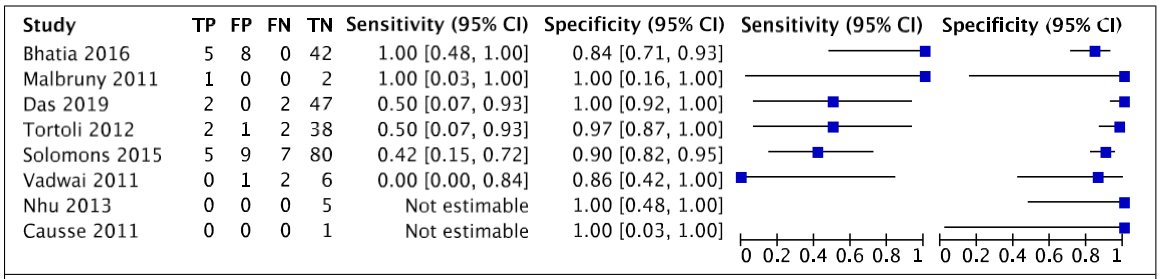


Figure 29. Q57 Forest plots of Xpert MTB/RIF sensitivity and specificity for tuberculosis meningitis using a cerebrospinal fluid specimen against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

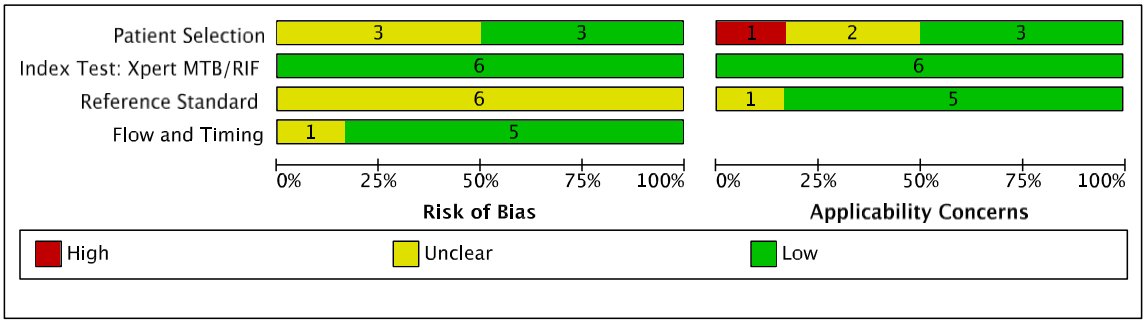


Figure 30. Risk of bias and applicability concerns graph for TB meningitis: review authors' judgements about each domain presented as percentages across included studies.

58. Should Xpert MTB/RIF be used to diagnose TB meningitis in CSF in children with signs and symptoms of TB meningitis, against a composite reference standard?

Xpert MTB/RIF sensitivity and specificity ranges were 25% to 38% and 100%, respectively (2 studies, 155 participants. We did not perform a meta-analysis owing to insufficient data.

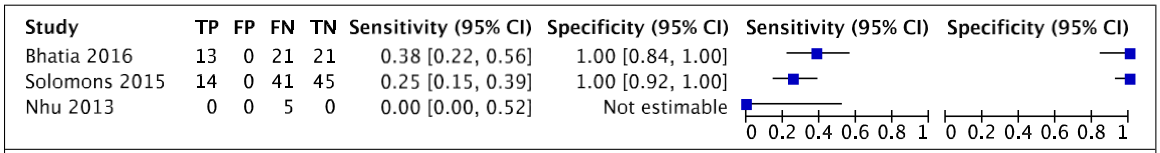


Figure 31. Q58 Forest plots of Xpert MTB/RIF sensitivity and specificity for tuberculosis meningitis using a cerebrospinal fluid specimen against a composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

59. Should Xpert MTB/RIF be used to diagnose lymph node TB in a lymph node specimen in children with signs and symptoms of lymph node TB, against a microbiological reference standard?

Xpert MTB/RIF pooled sensitivity and specificity were 90.4% (55.7 to 98.6) and 89.8% (71.5 to 96.8) (6 studies, 210 participants; very low-certainty evidence for sensitivity; low-certainty evidence for specificity

For a population of 1000 children where 100 people have lymph node TB on culture, 142 would be Xpert MTB/RIF-positive and 97 (68%) would not have lymph node TB (false-positives); 858 would be Xpert MTB/RIF-negative and 5 (1%) would have lymph node TB (false-negatives).

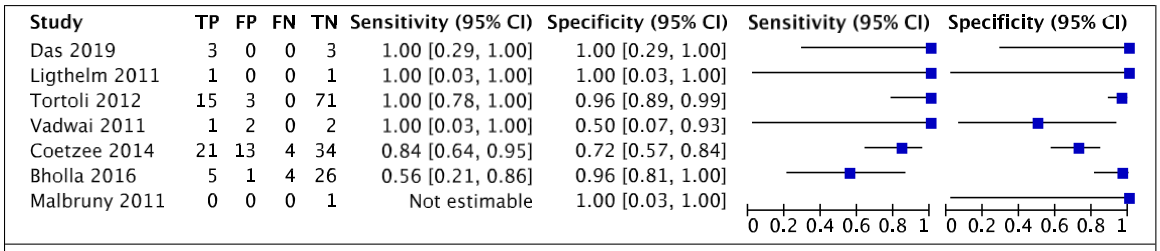


Figure 32. Q59 Forest plots of Xpert MTB/RIF sensitivity and specificity for lymph node tuberculosis using a lymph node specimen against a microbiologic reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

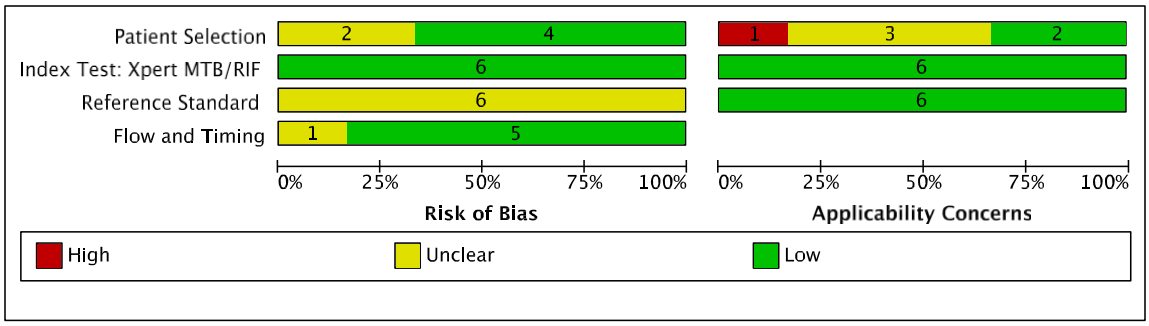


Figure 33. Risk of bias and applicability concerns graph for lymph node TB: review authors' judgements about each domain presented as percentages across included studies.

60. Should Xpert MTB/RIF be used to diagnose lymph node TB in a lymph node specimen in children with signs and symptoms of lymph node TB, against a composite reference standard?

Xpert MTB/RIF sensitivity and specificity ranges were 18% to 100% and 78% to 100%, respectively (2 studies, 105 participants). We did not perform a meta-analysis owing to insufficient data.

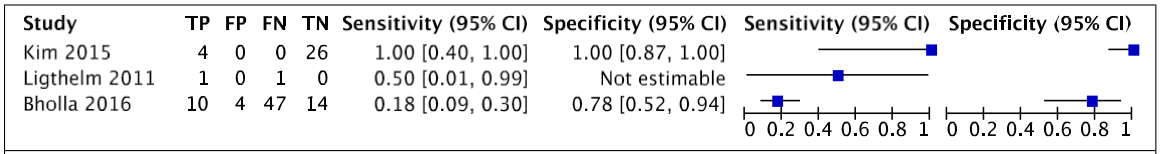


Figure 34. Q60 Forest plots of Xpert MTB/RIF sensitivity and specificity for lymph node tuberculosis using a lymph node specimen against a composite reference standard. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

Table 2. PICO 4 summary: diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for lymph node TB and TB meningitis

PICO, sub-question	Test, analysis group	Reference Standard	Studies	Number of children (cases)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive predictive value % (95% CI)	Negative predictive value % (95% CI)
57	Xpert MTB/RIF, CSF	MRS	6	262 (28)	54.0 (27.8 to 78.2)	93.8 (84.5 to 97.6)	49.1 (26.8 to 71.7)	94.8 (91.1 to 97.1)
58*	Xpert MTB/RIF, CSF	CRS	2	155 (89)	–	–	–	–
59	Xpert MTB/RIF, LN	MRS	6	210 (54)	90.4 (55.7 to 98.6)	89.8 (71.5 to 96.8)	49.6 (23.7 to 75.7)	98.8 (93.1 to 99.8)
60*	Xpert MTB/RIF, LN	CRS	2	105 (61)	–	–	–	–

*Meta-analysis was not possible due to limited data.

PICO 5: Detection of TB using multiple Xpert tests in Children
To evaluate multiple Xpert MTB/RIF (6 studies) or Ultra (2 studies) tests compared with a single test, we included studies that evaluated the diagnostic accuracy of single and multiple Xpert MTB/RIF or Ultra tests in children with signs or symptoms of pulmonary TB. There were 5 (1935 participants) studies that evaluated multiple Xpert MTB/RIF sputum specimens compared to one sputum specimen. The remaining analyses for other specimen types were limited to one or two studies.

Of the seven total studies, all had a low risk of bias for the domains on patient selection, reference standard and participant flow and timing. Two studies (29%) had high applicability concerns, and one (14%) unclear applicability concern for patient selection. All studies had low concern for the applicability of the reference standard.

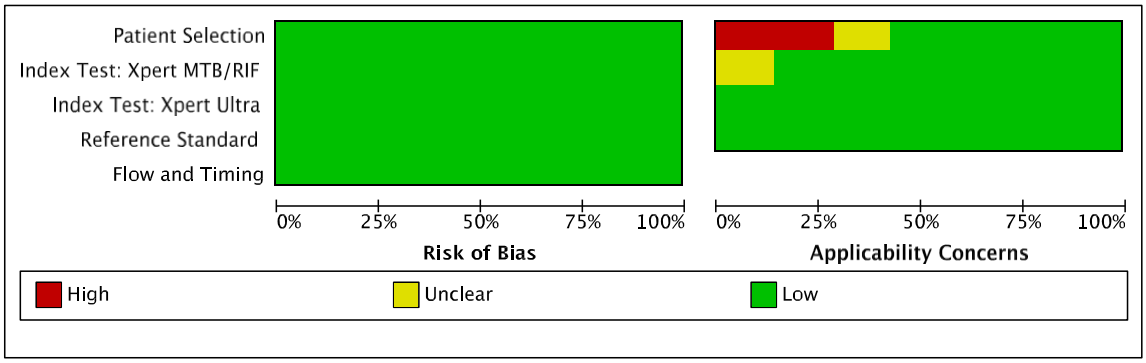


Figure 35. Risk of bias and applicability concerns graph for multiple Xpert studies: review authors' judgements about each domain presented as percentages across included studies.

PICO 5 subquestions

61. Should one Xpert MTB/RIF vs. more than one Xpert MTB/RIF be used to diagnose pulmonary TB in sputum in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?
The difference in the pooled sensitivity and specificity of Xpert MTB/RIF using multiple sputum specimens compared to one sputum specimen were 12.8% (-6.78 to 32.3) and -0.34% (-1.09 to 0.41) (5 studies, 1925 participants multiple Xpert MTB/RIF tests and 1939 participants one Xpert MTB/RIF test); low-certainty evidence for sensitivity; high-certainty evidence for specificity).

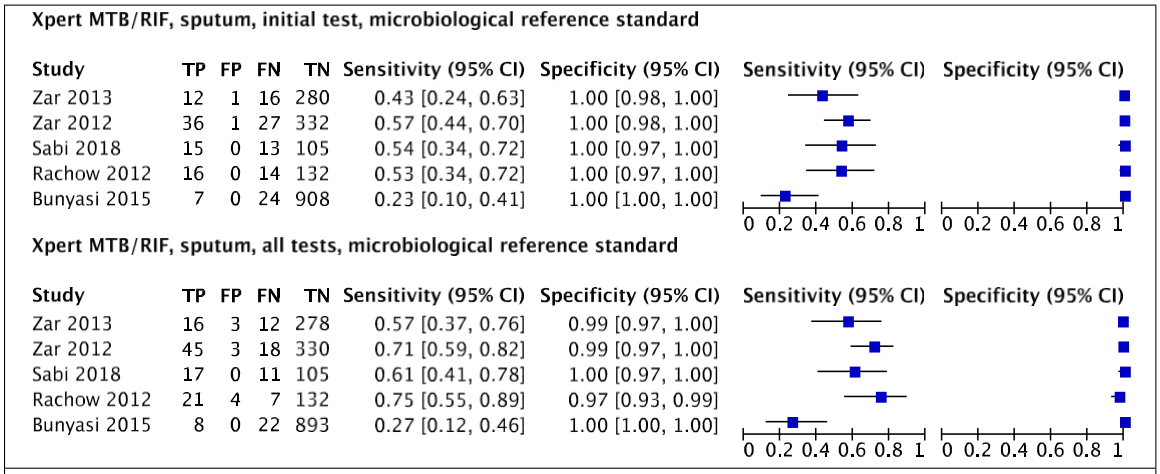


Figure 36. Q61 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum with one test against multiple tests. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

62. Should one Xpert Ultra vs. more than one Xpert Ultra be used to diagnose pulmonary TB in sputum in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

The difference in sensitivity and specificity for Xpert MTB/RIF using multiple sputum specimens compared to one sputum specimen were 10.7% (-13.2 to 34.6) and -1.87% (-4.44 to 0.70) (1 study, 135 participants; low-certainty evidence).

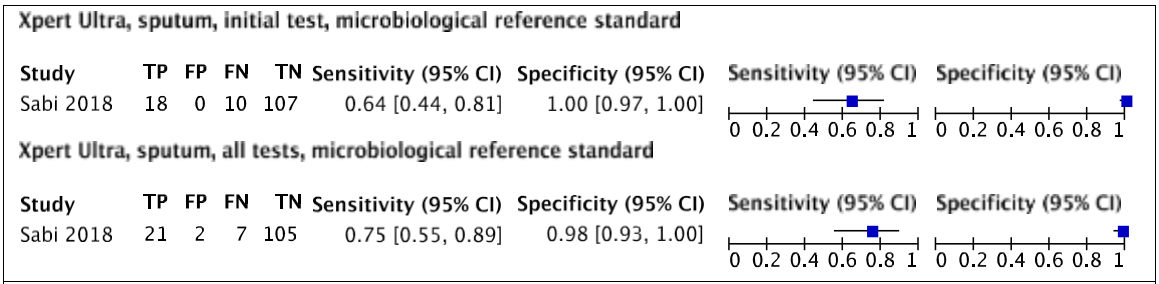


Figure 37. Q62 Forest plots of Xpert Ultra sensitivity and specificity for pulmonary tuberculosis using sputum with one test against multiple tests. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

63. Should one Xpert MTB/RIF vs. more than one Xpert MTB/RIF be used to diagnose pulmonary TB in a gastric specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

The difference in sensitivity and specificity for Xpert MTB/RIF using multiple gastric specimens compared to one gastric specimen were 13.2% (-4.64 to 31.1) and -0.45% (-0.99 to 0.09)) (1 study, 921 participants multiple Xpert tests and 935 participants one Xpert test; low-certainty evidence).

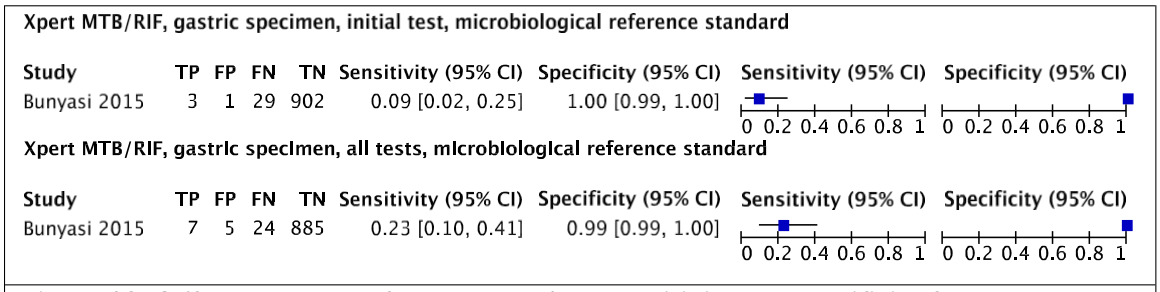


Figure 38. Q63 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using sputum with one test against multiple tests. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

64. Should one Xpert MTB/RIF vs. more than one Xpert MTB/RIF be used to diagnose pulmonary TB in a nasopharyngeal specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

The difference in pooled sensitivity and specificity for Xpert MTB/RIF using multiple nasopharyngeal specimens compared to one nasopharyngeal specimen were 13.5% (-9.50 to 36.5) and -0.49% (-1.63 to 0.66) (2 studies, 705 participants; very low-certainty evidence for sensitivity; moderate-certainty evidence for specificity).

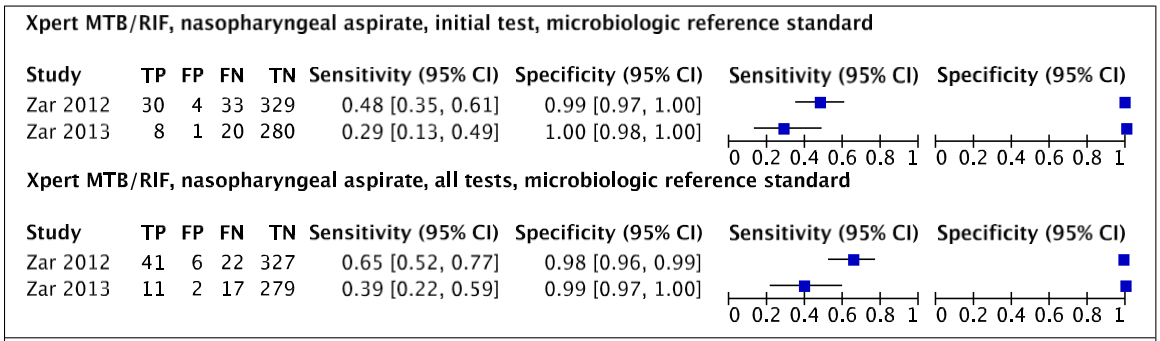


Figure 39. Q64 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using nasopharyngeal aspirates with one test against multiple tests. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

65. Should one Xpert Ultra vs. more than one Xpert Ultra be used to diagnose pulmonary TB in a nasopharyngeal specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

The difference in sensitivity and specificity for Xpert u MTB/RIF sing multiple nasopharyngeal specimens compared to one nasopharyngeal specimen were 16.7% (-11.1 to 44.5) and -1.89% (-6.34 to 2.57 (1 study, 130 participants; very low-certainty evidence for sensitivity; low-certainty evidence for specificity).

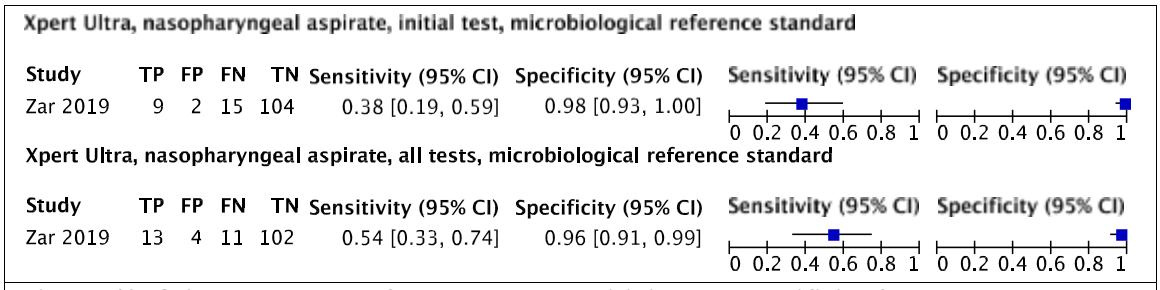


Figure 40. Q65 Forest plots of Xpert Ultra sensitivity and specificity for pulmonary tuberculosis using nasopharyngeal aspirates with one test against multiple tests. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

66. Should one Xpert MTB/RIF vs. more than one Xpert MTB/RIF be used to diagnose pulmonary TB in a stool specimen in children with signs and symptoms of pulmonary TB, against a microbiological reference standard?

The difference in sensitivity and specificity for Xpert MTB/RIF using multiple stool specimens compared to one stool specimen were 10.3% (-20.8 to 41.4) and 0.02% (-1.21 to 1.25 (1 study, 247 participants multiple Xpert MTB/RIF tests, 236 one Xpert MTB/RIF test; low-certainty evidence for sensitivity).

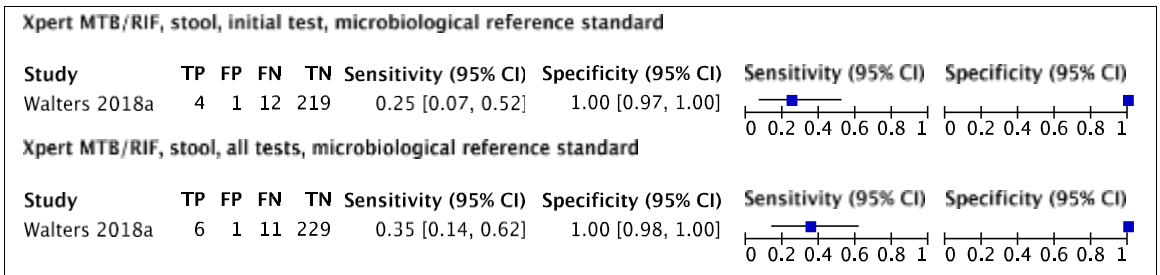


Figure 41. Q66 Forest plots of Xpert MTB/RIF sensitivity and specificity for pulmonary tuberculosis using stool with one test against multiple tests. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

Table 3. PICO 5 summary: diagnostic accuracy of Xpert MTB/RIF and Xpert Ultra for pulmonary TB in children, comparison of repeated testing versus first test

PICO, sub-question	Test, analysis group	Reference Standard	Studies	Number of children (cases)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive predictive value % (95% CI)	Negative predictive value % (95% CI)
61	More than one Xpert MTB/RIF, sputum	MRS	5	1925 (177)	59.1 (43.0 to 73.4)	99.5 (97.7 to 99.9)	93.5 (74.1 to 98.6)	95.6 (93.8 to 97.0)
61	One Xpert MTB/RIF, sputum	MRS	5	1939 (180)	46.3 (35.0 to 57.9)	99.9 (99.5 to 100)	97.8 (91.9 to 99.5)	94.3 (93.1 to 95.4)
	Absolute difference				12.8 (−6.78 to 32.3), P=0.20	−0.34 (−1.09 to 0.41), P=0.37		
62	More than one Xpert Ultra, sputum	MRS	1	135 (28)	75.0 (55.1 to 89.3)	98.1 (93.4 to 99.8)	81.7 (52.6 to 94.7)	97.2 (94.9 to 98.6)
62	One Xpert Ultra, sputum	MRS	1	135 (28)	64.3 (44.1 to 81.4)	100 (96.6 to 100)	Not estimable*	96.2 (93.8 to 97.6)
	Absolute difference				10.7 (−13.2 to 34.6), P=0.38	−1.87 (−4.44 to 0.70), P=0.16		
63	More than one Xpert MTB/RIF, gastric specimen	MRS	1	921 (31)	22.6 (9.59 to 41.1)	99.4 (98.7 to 99.8)	81.7 (60.0 to 93.0)	92.0 (90.5 to 93.4)
64	One Xpert MTB/RIF, gastric specimen	MRS	1	935 (32)	9.38 (1.98 to 25.0)	99.9 (99.4 to 100)	90.4 (50.1 to 98.9)	90.9 (89.9 to 91.7)
	Absolute difference				13.2 (−4.64 to 31.1), P=0.15	−0.45 (−0.99 to 0.09), P=0.10		
65	More than one Xpert MTB/RIF,	MRS	2	705 (91)	54.2 (36.1 to 71.3)	98.7 (97.4 to 99.3)	82.2 (69.6 to 90.3)	09.7 (96.7 to 95.1)

PICO, sub-question	Test, analysis group	Reference Standard	Studies	Number of children (cases)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive predictive value % (95% CI)	Negative predictive value % (95% CI)
	nasopharyngeal specimen							
65	One Xpert MTB/RIF, nasopharyngeal specimen	MRS	2	705 (91)	40.7 (27.9 to 54.9)	99.2 (98.1 to 99.7)	84.7 (69.8 to 93.0)	07.0 (92.3 to 95.0)
	Absolute difference				13.5 (−9.50 to 36.5), P=0.25	−0.49 (−1.63 to 0.66), P=0.40		
66	More than one Xpert Ultra, nasopharyngeal specimen	MRS	1	130 (24)	54.2 (32.8 to 74.4)	96.2 (90.6 to 99.0)	61.5 (36.3 to 81.7)	94.9 (92.4 to 96.7)
66	One Xpert Ultra, nasopharyngeal specimen	MRS	1	130 (24)	37.5 (18.8 to 59.4)	98.1 (93.4 to 99.8)	68.9 (33.8 to 90.5)	93.4 (91.2 to 95.0)
	Absolute difference				16.7 (−11.1 to 44.5), P=0.25	−1.89 (−6.34 to 2.57), P=0.41		
67	More than one Xpert MTB/RIF, stool specimen	MRS	1	247 (17)	35.3 (14.2 to 61.7)	99.6 (97.6 to 100)	90.0 (53.6 to 98.6)	93.3 (90.7 to 95.1)
67	One Xpert MTB/RIF, stool specimen	MRS	1	236 (16)	25.0 (7.27 to 52.4)	99.5 (97.5 to 100)	85.9 (42.0 to 98.1)	92.3 (90.0 to 94.0)
	Absolute difference				10.3 (−20.8 to 41.4), P=0.52	0.02 (−1.21 to 1.25), P=0.97		

*Not estimable because the predictive values were calculated using Bayes equation and likelihood ratios derived from the meta-analysis of sensitivity and specificity. Since specificity =100%, the false positive rate (FPR) is zero and so the positive likelihood ratio (sensitivity/FPR) cannot be computed due to division by zero

SUMMARY OF MAIN RESULTS

Xpert MTB/RIF pooled sensitivity (defined by culture) for the diagnosis of pulmonary TB
Sputum 64.6% (95% CI 55.3 to 72.9)
NPA 45.7% (95% CI 27.6 to 65.1)
Gastric specimen 73.0% (95%CI 52.9 to 86.7)
Stool 61.5% (95% CI 44.1 to 76.4)
Xpert MTB/RIF pooled specificity ranged between 98.1 and 100% for all specimen types used to diagnose pulmonary TB.

Xpert MTB/RIF pooled sensitivity defined by a composite reference standard for the diagnosis of pulmonary TB
Sputum 19.7% (95% CI 12.1 to 30.4)
Gastric specimen 31.7% (95% CI 20.2 to 46.0)
Stool 16.3% (95% CI 8.43 to 29.2)
Xpert MTB/RIF pooled specificity ranged between 99.7 and 100% for all specimen types used to diagnose pulmonary TB

Xpert Ultra pooled sensitivity defined by culture for the diagnosis of pulmonary TB
Sputum 72.8% (95% CI 64.7 to 79.6)
NPA 45.7% (95% CI 28.9 to 63.3)
Xpert Ultra pooled specificity was 97.5% for both specimen types used to diagnose pulmonary TB

Xpert Ultra pooled sensitivity defined by a composite reference standard for the diagnosis of pulmonary TB
Sputum 23.5% (95% CI 20.0 to 27.4)
Xpert Ultra pooled specificity was 99.2 (95% CI 96.9 to 99.8)

Xpert MTB/RIF pooled sensitivity defined by a microbiologic reference standard for the diagnosis of TB meningitis
CSF: Sensitivity 54.0% (95% CI 27.8 to 78.2)
CSF: Specificity 93.8% (95% CI 84.5 to 97.6)

Xpert MTB/RIF pooled sensitivity defined by a microbiologic reference standard for the diagnosis of lymph node TB
Lymph node: Sensitivity 90.4% (95% CI 55.7 to 98.6)
Lymph node: Specificity 89.8% (95% CI 71.5 to 96.8)

Multiple Xpert MTB/RIF or Ultra tests sensitivity as compared to one Xpert test for the diagnosis of pulmonary TB
Xpert MTB/RIF Sputum: 12.8% (95%CI −6.78 to 32.3), P=0.20
Xpert Ultra Sputum: 10.7% (95%CI −13.2 to 34.6), P=0.38
Xpert MTB/RIF Gastric Specimen: 13.2% (95% CI -4.64 to 31.1), P=0.15
Xpert MTB/RIF Nasopharyngeal Aspirate: 13.5% (95%CI −9.50 to 36.5), P=0.25
Xpert Ultra Nasopharyngeal Aspirate: 16.7% (95%CI −11.1 to 44.5), P=0.25
Xpert MTB/RIF Stool: 10.3% (95%CI −20.8 to 41.4), P=0.52
AUTHORS' CONCLUSIONS
Differences between sensitivity observed in this review may be in part attributable to differences in clinical setting (more commonly inpatient for gastric specimen collection) and differences in the quality of the reference standard.

The pooled sensitivity of Xpert Ultra was higher than Xpert MTB/RIF on sputum, but was unchanged for nasopharyngeal specimens; specificity was similar for sputum and nasopharyngeal aspirates (indirect comparisons). Xpert Ultra may detect an increased proportion of paucibacillary TB in children.

Xpert MTB/RIF sensitivity (defined by culture) was higher for multiple specimens compared with a single specimen and similar across specimen types. Multiple Xpert MTB/RIF or Ultra specimens of the same type, likely increases test sensitivity (defined by culture).

IMPLICATIONS FOR RESEARCH

There are numerous areas where additional research on the diagnostic accuracy of Xpert Ultra in children is necessary. Studies are needed that evaluate the diagnostic accuracy of Xpert Ultra in gastric or stool specimens for pulmonary TB and extrapulmonary TB in children. Establishing the diagnostic accuracy of Xpert Ultra on extrapulmonary specimens is an urgent need particularly given the encouraging data on cerebrospinal fluid in adults.

Xpert sensitivity increases with multiple as compared to one specimen in a small number of studies; however, there is a need for additional studies evaluating the combinatorial benefit of multiple specimen types. There were limited data suggesting that the combination of non-invasive specimens performs comparably with traditional gastric specimens or induced sputum specimens.

Additional operational and qualitative research is needed to determine the best approach to less-invasive specimen collection. Implementation studies on a method of suction for nasopharyngeal aspiration that is appropriate for low-skill or low-resource environments are needed. Extensive operational research is needed surrounding the use of stool as a diagnostic specimen in terms of integration into normal diagnostic clinical pathways, definition of laboratory protocols that successfully balance ease of implementation and diagnostic performance, and the impact of stool testing on patient important outcomes. There is a dearth of qualitative research identifying child and family preferences for and acceptability of comparative diagnostic approaches.

It is important to understand the diagnostic accuracy of Xpert MTB/RIF and Ultra in the context of the composite reference standard, where in most circumstances Xpert MTB/RIF and Ultra identify less than 30% of cases. More research continues to be needed to identify an improved reference standard that accurately defines TB disease in children. Sensitivity of all available diagnostics are sub-optimal and highlight the continued urgency to develop new tools that correctly diagnose a higher proportion of child TB cases. Ideally, the new tools will be rapid, affordable, feasible, and acceptable to children and their parents.

ACKNOWLEDGEMENTS

We thank Vittoria Lutje for developing the search strategy. We thank Aakshi Kaira and Claudia Denkinger for provision of a large dataset. We would like to thank Emily MacLean for providing data from her teams prior systematic review of Xpert on stool. Mikashmi Kholi provided valuable input on organizing the systematic review data extraction.

The Academic Editors are Professor Gerry Davies (CIDG) and Dr Danielle van der Windt (DTA Group).

CONTRIBUTIONS OF AUTHORS

AK, AMM, KRS, YT and LGF wrote the initial draft of the protocol. AD, AM, and AK designed the abstraction forms. AD, AM, AK, and LGF drafted the data analysis sections with input from KRS and YT. KRS drafted the QUADAS-2 section. KRS, YT, AM, and AD contributed methodological advice. All authors (AK, LGF, KRS, ME, YT, RV, AD, AM) provided input for the protocol.

DECLARATIONS OF INTEREST

AK has conducted prior primary research on tuberculosis diagnostics. He has an agreement with Cepheid for the supply of discounted Xpert cartridges to support a StopTB Partnership TB Reach case finding project. AK has received travel support to attend WHO guideline meetings.

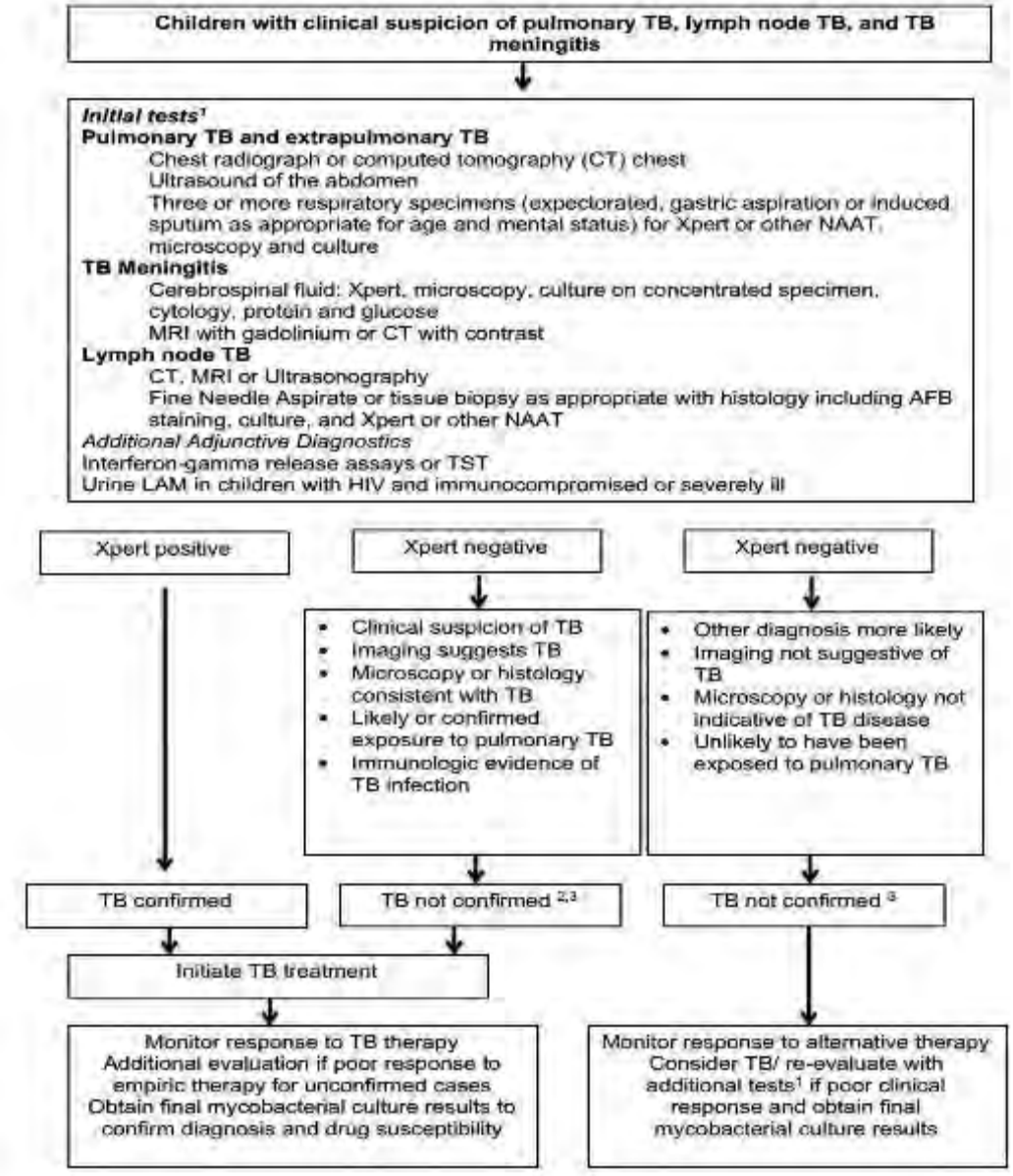
LGF has no known conflicts of interest.

YT has no known conflicts of interest.

ME has no known conflicts of interest.
RV has no known conflicts of interest.
KRS has received financial support for the preparation of systematic reviews and educational materials, consultancy fees from FIND (for the preparation of systematic reviews), honoraria, and travel support to attend WHO guideline meetings.
AD has conducted prior primary research on tuberculosis diagnostics, and has no known conflicts of interest.
AM has conducted prior primary research on tuberculosis diagnostics, and has received travel support to attend WHO guideline meetings.

APPENDICES

Appendix A. Clinical Pathway



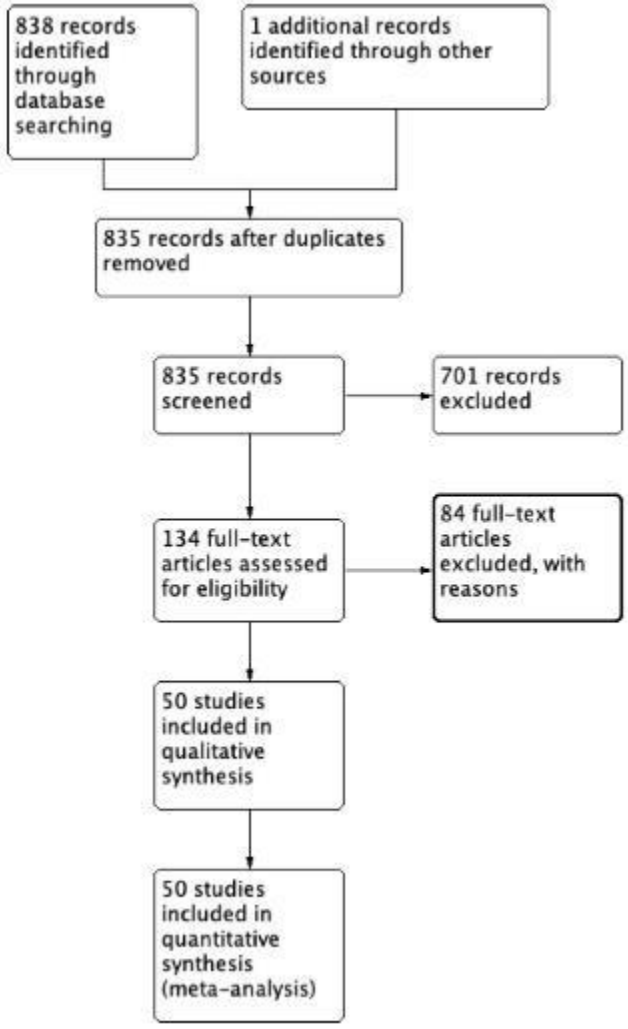
Abbreviations: AFB: acid-fast bacilli; CT: computed tomography; LAM: mycobacterial lipoarabinomannan antigen; MRI: magnetic resonance imaging; NAAT: nucleic acid amplification test; TB: tuberculosis; TST: tuberculin skin test. **The Clinical Pathway.** Clinical suspicion of tuberculosis includes persistent cough, fever, weight loss or failure to thrive, lymphadenitis, irritability, lethargy, headache, vomiting or neurological symptoms, history of possible or confirmed exposure to *M tuberculosis*, increased risk for tuberculosis disease due to immunocompromising conditions.

[1]Availability of investigations and tests may be different in high- and low-resource settings and may influence the approach to the diagnosis of child tuberculosis.

[2]Non-microbiological confirmation of *M tuberculosis* does not exclude tuberculosis disease in children, therefore initiation of treatment should be considered empirically if other clinical indications are present.

[3]Mycobacterial culture results are rarely timely to aid the decision to initiate treatment but can confirm or refute clinical decision-making if positive.

Appendix B. PRISMA Diagram and Reasons for Exclusion

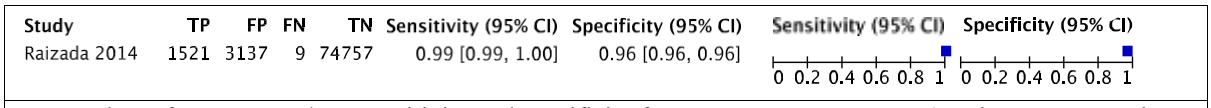


Exclusion reasons	Number of studies
Unable to separate paediatric data from adult data	20
Adult population	16
Not a diagnostic accuracy study	23
Case-control	3
Incorrect index test (index test not studied)	6
Unable to extract data by sample type	1
Insufficient data	5
Inappropriate reference standard	2
Index text not studied	1
Study involved screening for clinical tuberculosis prior to enrolment	2
Duplicate data	3
A clinical diagnosis of tuberculosis was established at enrolment	2
Total	84

Appendix C. Xpert MTB/RIF diagnostic accuracy with a references standard of smear

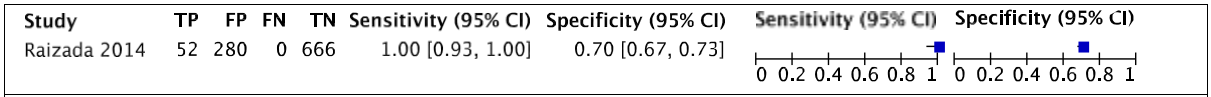
An analysis of a large dataset from India evaluating Xpert against a reference standard of smear demonstrated clearly different diagnostic accuracy than comparisons using a microbiologic reference standard of culture. The results for pulmonary TB, lymph node TB and TB meningitis are described below.

Xpert MTB/RIF sensitivity and specificity against smear as a reference standard for the diagnosis of pulmonary TB were 99% (0.99 to 1.00) and 96% (0.96 to 0.96), (1 study, 79,424 participants; grade assessment not performed).



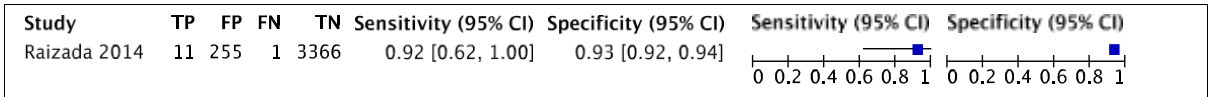
Forest plots of Xpert MTB/RIF sensitivity and specificity for **pulmonary tuberculosis** using sputum against a microbiologic reference standard of smear. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

Xpert MTB/RIF sensitivity and specificity against smear as a reference standard for the diagnosis of lymph node TB were 100% (0.93 to 1.00) and 70% (0.67 to 0.73), (1 study, 998 participants; grade assessment not performed).



Forest plots of Xpert MTB/RIF sensitivity and specificity for **lymph node tuberculosis** using lymph node specimens against a microbiologic reference standard of smear. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

Xpert MTB/RIF sensitivity and specificity against smear as a reference standard for the diagnosis of TB meningitis were 92% (0.62 to 1.00) and 93% (0.92 to 0.94), (1 study, 3633 participants; grade assessment not performed).



Forest plots of Xpert MTB/RIF sensitivity and specificity for **TB meningitis** using CSF against a microbiologic reference standard of smear. The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

Web Annex D.5. Systematic literature review of economic evidence for molecular assays intended as initial tests for the diagnosis of pulmonary and extrapulmonary TB in adults and children

Produced in Preparation for the WHO guideline development group meeting Dec 3-6, 2019

Draft Report v2
November, 2019

Alice Zwerling, Olivia D'Silva, Emily Thompson
School of Epidemiology & Public Health
University of Ottawa, Ottawa, Canada
azwerlin@uottawa.ca

Executive Summary

We carried out a systematic review of economic evaluations on molecular based tests GeneXpert MTB/RIF (Xpert) including the novel Xpert Ultra as well as the novel Molbio Truenat MTB/RIF test for the diagnosis of active TB. The objective of this review was to summarize current economic evidence and further understand the costs, cost-effectiveness and affordability of these molecular tests for TB diagnosis. We identified 28 studies meeting our inclusion criteria and addressing one of the PICO questions of interest. Only on study assessing cost-effectiveness of Truenat, and no studies were identified assessing cost-effectiveness of Xpert Ultra. Included studies were primarily assessing Xpert in African outpatient settings, but also among outpatients in India and Brazil. Four included studies were conducted among hospitalized patients in Germany, China (Hong Kong Special Administrative Region) and the USA, 2 screening studies focused on PLHIV in Mozambique and South Africa and one among the elderly in China (Hong Kong SAR) and among prisoners in former Soviet Union countries. One study specifically assessing extrapulmonary TB was conducted in Beijing, China, and repeated Xperts were evaluated in two studies from South Africa among PLHIV, one among hospitalized patients from the US and one study among outpatients in China.

Studies employed a variety of different modelling approaches, populations and settings. Variations in costing, effectiveness and epidemiological parameters were present across included studies making direct comparisons across studies challenging. Studies used both short-term diagnostic outcomes (additional cases diagnosed, RR-TB cases diagnosed) and long-term outcomes (years of life saved, DALYs averted etc.). There was variation in costing elements included across different analyses, both in terms of what was included in unit test costs (consumables and equipment only versus overhead, staffing, training etc.), and whether implementation costs were included for introducing novel diagnostic testing, and whether downstream costs associated with TB treatment, MDR-TB treatment and ART and HIV care were included.

While many studies demonstrated that Xpert may be cost-effective in diagnosing pulmonary TB, key implementation conditions and settings were highly influential on determining cost-effectiveness and must be considered when implementing Xpert. Cost-effectiveness of Xpert was shown to be improved among populations with higher TB prevalence, among PLHIV populations and where rates of empirical treatment were low. Cost-effectiveness of Xpert is highly dependant on a number of important factors including placement of Xpert machines (centralized facilities versus decentralization) test volume, underlying TB prevalence, level of empirical treatment and pre-treatment loss to follow-up. Only one study assessing cost-effectiveness of Molbio's Truenat MTB/RIF was identified, while this study suggests Truenat is likely cost-effective if implemented at the POC in India, it relies on several important modelling assumptions including improved linkage to care and increased treatment initiation.

Caution should be used when generalizing cost-effectiveness and economic evaluations across settings. Local implementation conditions and settings should be considered and local implementation studies may be helpful to assess likely impact on case-finding, long-term outcomes and cost-effectiveness.

BACKGROUND

The Xpert™ MTB/RIF assay (Cephid, Sunnyvale, Ca. USA) is an automated, cartridge-based nucleic acid amplification test for the detection of *Mycobacterium tuberculosis* employing the GeneXpert™ multi-disease platform. The Xpert assay can be performed on sputum samples (processed or unprocessed) as well as selected extrapulmonary specimens. Xpert provides rapid turn-around time to results, typically within 2 hours and also provides identification of rifampicin resistance. In 2010, WHO endorsed use of this diagnostic test for TB and subsequent policy updates have been issued in 2013 and 2016.

Truenat (Molbio Diagnostics/Bigtec Labs, Goa/Bengaluru, India) is a novel molecular assay that offers rapid detection of tuberculosis (TB) and rifampicin-resistance using a battery powered platform which may be potentially useful in peripheral healthcare settings or for point-of-care testing.

In preparation for the guideline development group meeting "Molecular assays intended as initial tests for the diagnosis of pulmonary and extrapulmonary TB in adults and children. Policy Update" scheduled for 3-6 December, 2019, there is a need to summarize the current economic evidence on molecular tests: GeneXpert MTB/RIF/Ultra and the novel Molbio/TrueNat MTB/RIF including available cost evidence, cost-effectiveness, and affordability of these tests for the diagnosis of TB and Rifampicin resistance (RR).

Molecular tests have demonstrated potential to improve case-finding over standard approaches including sputum smear microscopy and offer rapid turn around time compared with culture, but likely also come with a higher unit test cost. Understanding the costs, cost-effectiveness and affordability of these molecular tests across different populations and settings individuals can provide important evidence for policy makers needing to make decisions around scale-up of Xpert and/or Truenat within national TB programmes.

Objective

To perform a systematic review of the published literature on economic evaluations on the molecular based tests GeneXpert MTB/RIF / Ultra and Molbio TrueNat MTB/RIF for the diagnosis of active TB. To summarize current economic evidence and further understand the costs, cost-effectiveness and affordability of these molecular tests for TB diagnosis. Affordability will be considered with respect to budget impact assessments performed in specific countries under given scenarios/conditions.

Research Questions

PICO 1 & 2: Among adults and children with signs and symptoms of pulmonary TB, seeking care at health care facilities what is the economic evidence surrounding the use of Xpert MTB/RIF / Ultra as an initial test for diagnosis of pulmonary TB and RR?

PICO 3 & 4: Among adults and children with signs and symptoms of EP TB, seeking care at health care facilities what is the economic evidence for Xpert MTB/RIF / Ultra to be used as an initial test for diagnosis of EP TB and RR?

PICO 5: Among people with signs and symptoms of pulmonary TB, seeking care at health care facilities what is the cost-effectiveness of repeated Xpert (Ultra) tests on subsequent samples for the diagnosis of pulmonary TB?

PICO 6: Among adults in a population-based TB disease prevalence survey with symptoms or chest X-ray abnormalities suggestive of pulmonary TB, what is the economic evidence for Xpert MTB/RIF / Ultra alone, or to be used to define the case of active TB disease[4]?

PICO 7: Among people being screened for pulmonary TB, what is the economic evidence for Xpert MTB/RIF/Ultra to be used alone to define TB and RR?

PICO 8: Among people with signs and symptoms of pulmonary TB, seeking care at health care facilities what is the economic evidence and cost-effectiveness of Molbio TrueNat MTB / Rif to be used as an initial test for diagnosis of pulmonary TB and RR?

METHODS

Search strategy & Data Sources

We performed a search of three online databases: EMBASE, Medline, Web of Science and Scopus for new studies published from January 1, 2007 through October 4, 2019 We reviewed citations of all eligible articles, guidelines and reviews for additional studies.

The search strategy used was modified to meet the criteria of each database but generally included the following terms and structure: (tuberculosis OR TB OR mycobacterium)
AND ((xpert* or genexpert* or cepheid* or molbio or truenat)) *AND* ("cost-benefit*" OR cost* OR economic* OR "cost effectiveness*" OR "cost-utility" OR "disability adjusted life year*" OR DALY OR "quality-adjusted life year*" OR QALY OR "cost benefit analysis" OR "cost effectiveness analysis" OR "quality of life" OR "utility").

Types of studies considered

Studies were included if they evaluated one of the three tests under investigation: Xpert MTB/RIF, Xpert MTB/RIF Ultra or Molbio TrueNat MTB/RIF for the detection of active TB disease and included an economic evaluation in the analysis. Our search terms as outlined above, were designed to broadly capture any economic evaluations or studies mentioning cost or disability-adjusted life years (DALYs) or quality-adjusted life years (QALYs) and was not limited to cost-effectiveness analyses. We considered studies eligible if they evaluated the use of either Xpert or Truenat as the initial diagnostic test. Eligible studies could use either primary or secondary data sources (i.e., published literature) for either economic or epidemiologic parameters and included studies using both hypothetical modelled populations and those based on diagnostic trials. Studies were excluded from full data extraction if there was no link to health outcomes such as incremental yield, mortality, or DALYs/QALYs. Only studies published in English were included. While only studies including cost-effectiveness analyses were included in the systematic review for full data extraction, studies focusing solely on test costs and/or implementation are included when relevant in the subsequent discussion.

Study selection

The study selection followed PRISMA guidelines (1,2). Potentially relevant studies were identified through electronic searches of the online databases as described above, and duplicates were removed. An initial abstract review of each study was completed independently by two reviewers (OD & ET);

articles were excluded if they did not evaluate one of our diagnostic tests, or if they were reviews, letters or opinion pieces (i.e. no original data), conference abstracts were also excluded. Full text review was then completed on remaining articles, and articles that met predetermined inclusion criteria were retained for the review.

Data extraction

Full texts of included studies including published supplemental material, were independently reviewed by two reviewers, with all disagreements resolved by discussion with a third reviewer.

The study design data elements extracted from each study included: the primary research question, country and setting, year of study, patient population, clinical setting, Xpert diagnostic scenarios, comparison diagnostic scenarios and reference scenarios, economic analysis perspective, analytic time horizon, type of economic evaluation, source of costing, primary outcome measure, type of model, types of sensitivity and uncertainty analyses performed and willingness-to-pay threshold.

Key model parameters were extracted and presented in tables, including epidemiologic parameters, diagnostic accuracy parameters and treatment and outcome parameters. Cost components and unit test costs were also extracted, along with key costing input parameters. Costs are presented in USD (United States Dollars).

RESULTS

Study Selection

A search of online databases as of October 4, 2019 returned a total of 1221 articles. Duplicates were excluded (n=612) for a total of 609 articles that were screened and reviewed for study eligibility. Articles were excluded if they were not relevant to our study question (n=469, not focused on TB, did not include Xpert or Truenat, or did not include an economic analysis). A full text review was performed on the remaining 140 articles, 112 studies were excluded for not meeting inclusion criteria outlined in the above methods section, leaving 28 articles eligible for inclusion in our review. Details on reasons for exclusion are available in the PRISMA flow diagram in figure 1.

Characteristics of included studies

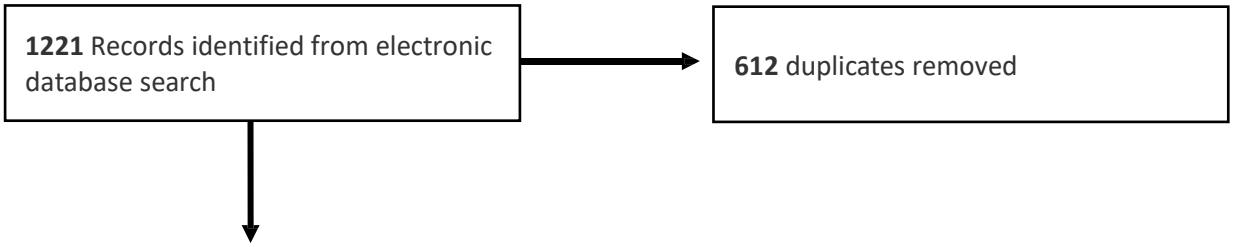
Study characteristics for included studies are presented in tabl 1. Among the 28 studies included in this review, 17 focus on economic evaluations of Xpert in sub-Saharan Africa (3–19), 10 with analyses based in South Africa (5–8,10,11,13,14,16,19), and three of which include multi country analyses (10,13,19). Economic analyses from African countries included Ethiopia, Tanzania, Botswana, Lesotho, Namibia, Swaziland, Mozambique, Zambia, Zimbabwe, Uganda and Malawi. Three studies assess cost-effectiveness in high income/low TB prevalence countries including Germany and the United States (20–22), 4 studies include analyses of cost-effectiveness of Xpert in India (19,23–25), 1 in China (26), 2 in China (Hong Kong SAR)(27,28), 1 in Brazil (29) and 1 from the former Soviet Union countries (30).

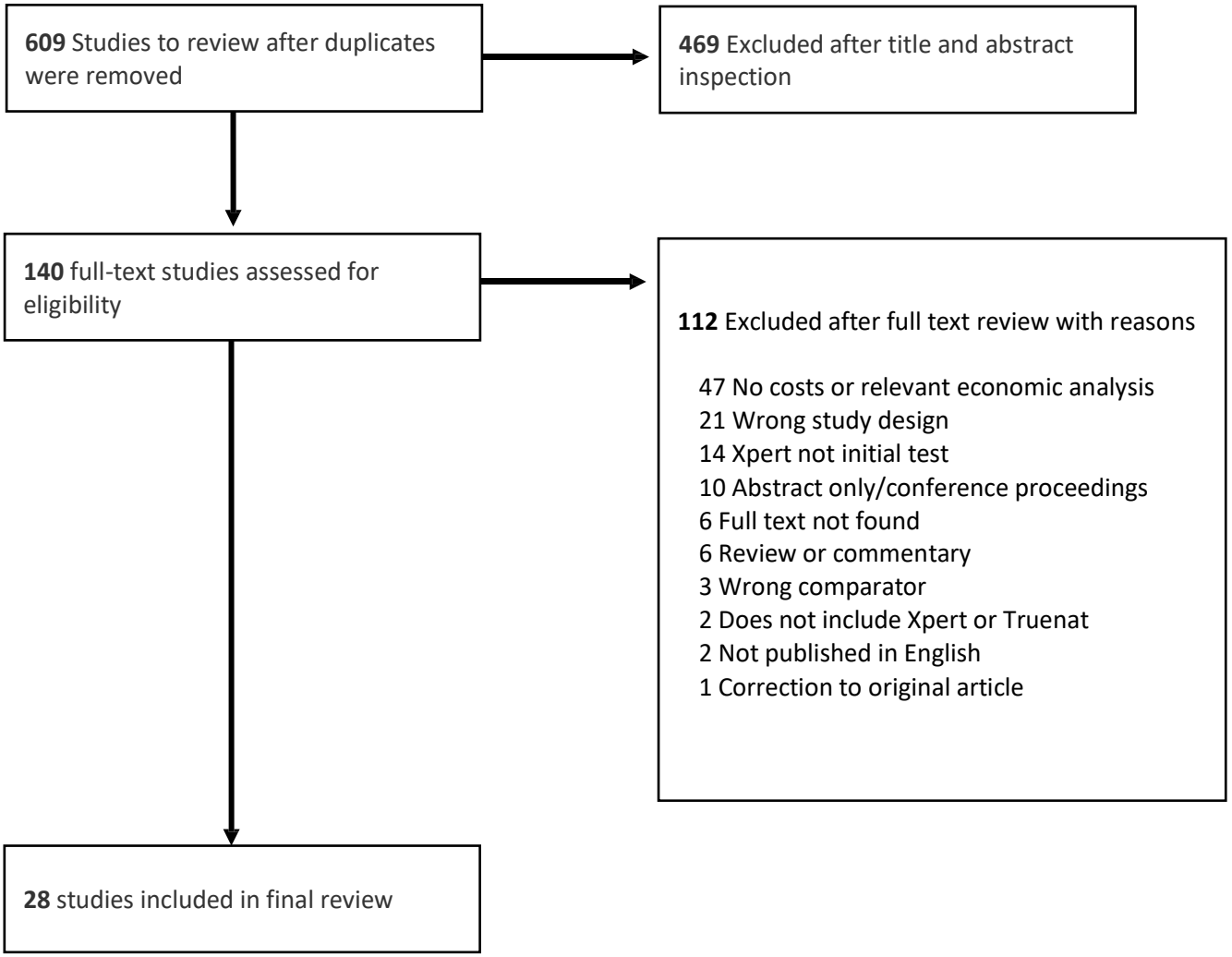
A large number of economic analyses have been published since the 2016 review and WHO policy update, including 16 studies published between 2016 and 2019 (4,6–8,12,13,15–17,20,21,24–27,29). Among the 28 included studies, 5 studies focused exclusively on people living with HIV (PLHIV) all from sub-Saharan Africa: Ethiopia, South Africa, Mozambique and Malawi (3–5,12,18). Four studies focus their cost-effectiveness evaluations on hospitalized patients: 2 from the United States (21,22),

1 from Germany (20) and 1 from China (Hong Kong SAR)(28). Fifteen studies assessed predominantly outpatient settings or used population level approaches. We identified only one study assessing cost-effectiveness of Xpert specifically in extrapulmonary samples (26). Regarding the evaluation of repeated Xpert we identified 2 studies focused solely in PLHIV (5,14), one study among hospitalized populations (21) and one study among outpatients (26). No studies identified in this review assessed the cost-effectiveness of Xpert for use as an initial diagnostic test in a population-based prevalence survey. Cost-effectiveness of screening in select populations was assessed in two studies focused on screening among PLHIV (5,12) and by Li et al in China (Hong Kong SAR) with their evaluation of Xpert among patients with symptoms of TB at admission to residential care homes for the elderly (27), and Winetsky et al who evaluated Xpert for screening inmates in prisons across former Soviet Union countries (30). Finally we identified one that directly addressed the cost-effectiveness of Molbio Truenat MTB/RIF study (24).

We had restricted our inclusion criteria to only include studies where at least one diagnostic algorithm under investigation included Xpert as the initial diagnostic test. Some studies focused on implementing Xpert in centralized testing facilities while others evaluated Xpert as a decentralized or point-of-care approach. A variety of reference strategies were compared but most typically included sputum smear microscopy (SSM) and/or culture. Most studies took a health care system perspective with just two (16,23) employing a societal perspective and including patient costs. Studies employed a variety of modelling approaches, epidemiological and costing parameters and implementation approaches making comparisons across different studies challenging. Studies included a mixture of modelling approaches, from decision analysis, and discrete event simulation to transmission model while several studies were based on trial or operational data without modelling components. Primary outcomes across studies included incremental cost/DALY averted or incremental cost/QALY gained, incremental cost /year of life saved, cost per accurate TB diagnosis or Rifampicin resistant TB case identified, and with the exception of Schnippel and Diel, costs were presented in USD and studies employed a mix of published literature and empirical costing. Several studies also included budget impact assessments, details of which have been summarized below in relevant sections. Willingness to pay thresholds varied greatly across studies to determine cost-effectiveness, but relevant country's GDP per capita was most commonly used. Cost components of unit test costs are presented in table 2, with checkmarks indicating when authors explicitly mentioned elements were included in unit test cost calculations.

Figure 1. PRISMA flow diagram showing study inclusion and exclusion for systematic review of Xpert and Truenat studies.





PICO 1 & 2: Among adults and children with signs and symptoms of pulmonary TB, seeking care at health care facilities should Xpert MTB/RIF/ Ultra be used as an initial test for diagnosis of pulmonary TB and Rif resistance?

Cost-effectiveness of Xpert MTB/RIF Among People living with HIV with signs and symptoms of TB

We identified four studies assessing the use of Xpert MTB/RIF among PLHIV with signs and symptoms of TB (3–5,18). Studies were conducted in countries with high HIV prevalence including South Africa, Ethiopia and Malawi. All reported Xpert would likely be cost-effective in these populations but to varying degrees and conditions of implementation. No studies assessed children specifically among these studies.

Abimbola et al assessed Xpert as a replacement for SSM followed by CXR if smear negative in a population of persons with advanced HIV initiating ART in South African HIV clinics and found Xpert dominated current practice as the least costly and most effective algorithm at reducing early mortality compared with SSM and CXR (3). Limitations of this analysis include not accounting for empirical treatment or implementation costs.

Andrews et al also assessed a population of PLHIV initiating ART in South Africa, and compared several diagnostic algorithms both in patients with signs and symptoms of TB and among all patients regardless of symptoms (5). Employing 2 Xperts for all (regardless of symptoms) was more cost-effective and less costly, with a lower ICER (US$ 5100/YLS) compared with 2 cultures or 2 Xperts among symptomatic persons or a single Xpert regardless of symptoms. This remained cost-effective (using the South African GDP as a willingness to pay threshold) unless TB prevalence fell below 7.5%. This analysis does allow for empirical treatment but did not account for costs associated with Xpert scale- up nor did they account for potential benefits accrued through ongoing transmission.

Zwerling et al. used costs and operational data from an RCT assessing point of care diagnostics including LED fluorescence microscopy and Xpert among people with a new HIV diagnosis and signs and symptoms of TB in rural Malawi (18). Authors found Xpert could be cost-effective compared to standard of care - SSM at the discretion of the physician - (using Malawian GDP) given a combination of high-test volume at point of care and TB prevalence (ICER US$ 298 /DALY averted (1000 samples/annually and 6% TB prevalence). Whereas in settings with low annual test volume, and/or low TB prevalence LED would be preferred over Xpert, ICER US$ 6606/DALY averted (50 samples tested annually and 1% TB prevalence).

In the most recent publication, Adelman et al assess Xpert among those with signs and symptoms of TB in Ethiopian HIV clinics, compared with standard approach of smear microscopy and clinical diagnosis among a population where 89% are already on ART (4). Authors found the Xpert algorithm to be highly cost-effective with an ICER of US$ 5/DALY averted at a 6% TB prevalence, assuming 15,000 Xpert tests performed annually. Authors did not account for scale-up costs, impacts to ongoing transmission, or patients unable to provide sputum or those lost to follow-up.

Cost-effectiveness of Xpert among hospitalized patients with signs and symptoms of TB

Four studies among hospitalized patients were identified, 2 from the USA (21,22), 1 from Germany (20) and 1 study from China (Hong Kong SAR)(28). All 4 studies concluded that replacement of SSM with Xpert would results in cost-savings driven largely from high hospitalization costs associated with respiratory isolation. No studies assessed children specifically among these studies.

Millman et al assessed using Xpert as a replacement for 2 SSM amongst hospitalized patients with signs and symptoms of TB in San Francisco, USA (22). Authors concluded Xpert strategy would result in cost-savings of US$ 2278/patient compared with SSM, cost savings were driven by reduction of costs associated with admission for respiratory isolation.

You et al also assessed the replacement of 2 SSM with Xpert among patients with signs and symptoms of TB hospitalized in China (Hong Kong SAR)(28). The Xpert strategy dominated and was cost-savings and more cost-effective compared with 2 SSM. This result was robust unless sensitivity of SSM was increased from 50% to over 74% where a SSM followed by Xpert only I those with negative smear results was preferred.

Diel et al performed a cost-benefit analysis in a German inpatient population to assess the replacing smear and culture with Xpert for diagnosis amongst patients with suspected TB (20). Authors found Xpert resulted in a cost-savings of 449.98 Euro/patient primarily driven by reduction of cots association with isolation in hospital. Authors note downstream costs such as expanding subsequent contact investigation to all patients as opposed to only smear positive patients would likely lead to increased costs not explicitly accounted for in this analysis.

Cowan et al assessed the use of Xpert amongst hospitalized patients suspected of pulmonary TB in Seattle, USA (21). Compared with SSM on either 2 or 3 sputum samples, 1 unconcentrated Xpert was cost savings and more effective thus dominating the standard of care. Cost savings were primarily driven by high costs of airborne infection isolation. The 2 Xpert strategy was also cost-savings when compared with 3 SSM, but resulted in a very high ICER of $2826682/accurate case diagnosed compared with 1 Xpert.

Cost-effectiveness of Xpert outpatients with signs and symptoms of TB

We identified 15 studies assessing cost-effectiveness of Xpert among persons presenting to primary health care facilities across South Africa, India, Uganda, Botswana, Lesotho, Namibia, Swaziland, Tanzania, Mozambique, Ethiopia, Zambia, Zimbabwe and Brazil (6–11,13,15–17,19,23–25,29). While early studies found Xpert would likely be cost-effective (albeit using a range of willingness to pay thresholds across different countries, several concerns around cost-effectiveness have been raised by subsequent analyses. Inclusion of downstream costs associated with MDR-TB and HIV treatment and care has been shown to lead to increased ICERs and increased total expenditures. Costs associated with scale-up of Xpert have been estimated to result in an important increase relative to existing TB and HIV programme budgets and in many countries may not be deemed affordable despite ICERs for Xpert approaches being under willingness to pay thresholds. Studies have highlighted the importance of implementation conditions, including existing standard of care, levels of empirical treatment, TB prevalence among presumptive patients being tested, and test volume as highly influential variables on cost-effectiveness results. Results from individual studies are summarized below. While some studies employd a population based approach no studies specifically addressed children.

Vassall et al (PLos Medicine 2011) was one of the first to assess the cost-effectiveness of replacement of SSM with Xpert among outpatients presenting with signs and symptoms of TB, and focused on three different countries: South Africa, India and Uganda (19). Authors found model increased case-finding and costs resulting in ICERs of US$ 68/DALY averted in India, US$ 138/DALY averted in South Africa and US$ 52/DALY averted in Uganda. Limitations of this early work included not accounting for ART costs or patient costs, empirical treatment or transmission.

Meyer-Rath used a population-level model to assess cost-effectiveness of Xpert as a replacement for the current SSM and culture-based approach in South Africa among patients with suspected TB (11). Authors concluded that at full scale, Xpert will increase the cost of national TB diagnosis and treatment programmes it will also increase the number of TB cases diagnosed per year by 69-71%, notably increasing proportion of patients diagnosed on the initial visit from 46% to 81%. Cost of TB diagnosis per patient tested increased by 55% to US$ 61. The incremental capital cost of Xpert scale-up was estimated to be 22 million with an incremental recurrent cost of US$ 287-316 million over 6 years.

Menzies et al employed a dynamic transmission model examining replacement of SSM with Xpert in 5 countries: Botswana, Lesotho, Namibia, South Africa and Swaziland (10). Authors found ICERs over a ten-year horizon for Xpert ranged from $792 in Swaziland to $1257 in Botswana, and study found ICERs continue to fall 20% over 20 years due to benefits accrued through reduced transmission.

ICERs were significantly higher than had been reported in previous CEA partially due to inclusion of downstream ART costs, which along with MDR-TB treatment costs account for a large proportion of total expenditures under the Xpert strategy.

Langley et al employed a linked operational and transmission modelling approach to assess cost-effectiveness of full roll- out of Xpert as a replacement for SSM among outpatients in Tanzania presenting with signs and symptoms of TB (9). Authors estimated an ICER of US$ 169/DALY averted for Xpert compared with SSM, which was deemed cost-effective compared to the Tanzania GDP (US$ 599), authors also estimated the additional cost of roll-out at US$ 36.9 million over 10 years, representing an increase of 25% compared to current TB and HIV programmes. These results were based on low levels of empirical treatment observed in Tanzania and authors note results may not be generalizable countries with high rates of empirical treatment. Similar to Menzies study above, this analysis does include transmission and costs of ART.

Suen et al also employed a dynamic transmission model this time in the Indian context to assess the cost-effectiveness of Xpert as a replacement for SSM and culture (when MDR-TB is suspected) (23). Suen was also one of just a few studies to take a societal versus health care perspective and Suen et al also sought to model care-seeking in both the public and private sectors in India, and were interested in assessing the cost-effectiveness of implementing nationwide public-private mix (PPM) assuming a cost of US$ 38 /person. In the evaluation of Xpert as the initial test to replace SSM authors found that while more individuals received accurate diagnoses and time to treatment initiation among MDR-TB patients was greatly reduced, this strategy was dominated by PPM strategies.

Jha et al performed a cost-effectiveness analysis looking at automated microscopy for outpatients with signs and symptoms of TB in South Africa but also assessed Xpert for all and manual SSM (8). Compared to manual SSM, Xpert for all resulted in an ICER of US$ 1720/additional true TB diagnosis. Test volume was an important factor in analyses as authors were interested in peripheral testing setting. Authors concluded that where resources are sufficient Xpert is the preferred strategy.

Wikman-Jorgensen et al employed a transmission markov model to assess the cost-effectiveness of Xpert or Microscopic observation drug-susceptibility assay (MODs) compared with SSM in rural Mozambique (17). Model assumes no empirical treatment and no MDR-TB and estimates and ICER of US$ 122.13/DALY averted for Xpert compared with SSM. Several parameters were influential on model results in sensitivity analyses specifically TB prevalence and risk of infection, therefore authors concluded that while Xpert was found to be cost-effective (using Mozambique GDP as WTP threshold) uncertainty was high.

Khaparde et al used a decision analysis model to assess the scale-up of Xpert as a replacement for SSM in the Indian context (25). Total costs associated with the Xpert for all approach increased by 46% compared with SSM primarily due to costs associated with second-line treatment of a higher number of rifampicin-resistant patients due to increased drug-resistant TB. Diagnostic costs for an estimated 7.64 million presumptive TB patients would account for 50% of the annual TB control budget in India. Mean total costs per DR-TB case initiated on treatment were lowest in the Xpert for all scenario.

Tesfaye et al assessed 8 different algorithms including an Xpert for all scenario compared with the standard of care: 3 SSM, among persons presenting to public health facilities in Ethiopia with signs and symptoms of TB (15). Full roll-out of Xpert is expected to produce the greatest patient level gains, and an ICER of US$ 370/DALY averted compared to SSM, and an additional health system cost of US$ 11.6 million USD over 10 years. Authors concluded that full roll out of Xpert was not affordable and recommended same day LED fluorescence microscopy as an alternative combined with targeted

Xpert. Targeted approaches may be more affordable but do not convey the same amount of health benefits.

Vassall et al (Lancet GH 2017) employed XTEND (a pragmatic trial) data which assessed the use of Xpert in lieu of SSM among people presenting to primary health facilities in South Africa with signs and symptoms or TB (16). This is one of just a few studies to employ a societal perspective and include patient costs. Authors found less than a 3% probability that Xpert is cost-effective in this population, primarily due to results from the XTEND trial showing no improvement in time to treatment initiation, no increase in number of people initiating ART and was not powered to examine time to initiation amongst MDR-TB patients. In most simulations, model results showed fewer DALYs averted as treatment initiations were actually lower in Xpert arm compared with SSM. Xpert roll-out was predicted to be cost-neutral where increased costs associated with Xpert equipment was mitigated by reduction in downstream costs further along the diagnostic cascade. Cost-effectiveness of Xpert was influenced by implementation and adherence to to TB and HIV diagnostic and treatment pathways and less influenced by Xpert's specific diagnostic performance.

Dunbar et al used an operational model in South Africa to assess the effects of increased case-finding (scale-up of Xpert) and corresponding decrease in proportion of TB cases among presumptive TB patients on laboratory costs per TB case diagnosed and per additional TB case diagnosed with Xpert versus SSM and culture (6). Authors varied TB prevalence and Xpert cartridge costs. Cost per additional TB case diagnosed in Xpert strategy compared with SSM and culture was US$ 986. Cost ranged from US$ 603 at 31% TB prevalence to US$ 9245 at 3%. Varying Xpert cartridge cost resulted in cost per additional TB case diagnosed via Xpert of $886 given a 10% cost reduction per cartridge to US$ 489 at a 50% cost reduction. I an ideal situation if TB prevalance among presumptive patients were 25- 31% and price of Xpert cartridges could be reduced by 50%, the cost per TB case diagnosed would range from US$ 50-59, comparable to cost per TB case diagnosed via SSM and culture ($48.77).

In a subsequent analysis Dunbar et al estimated the number and cost of rifampicin resistant TB cases identified using smear/culture and Xpert algorithms (7). The Xpert algorithm increased the number of RR-TB cases diagnosed from 603 with smear/culture to 1178 with Xpert. The cost per RR-TB case identifies increased from US$ 1781 with smear/culture to US$ 2063 with Xpert.

Pooran et al used data from the TB-NEAT trial to compare point of care (POC) SSM with POC Xpert among patients with signs and symptoms of TB presenting to primary health care facilities across four African countries (South Africa, Zambia, Zimbabwe and Tanzania) (13). As has been identified by other groups assessing point of care or implementation at peripheral centres, Xpert costs were sensitive to test volume, and cost for Xpert was greater at POC (US$ 28.03) compared with centralized lab/facility (US$ 23). Compared with SSM, Xpert at POC cost an additional US$ 1464 per treatment initiation and an additional US$ 1211 per completion. Probability of POC Xpert being cost-effective was 90% using a willingness to pay of $3820 per treatment completion. This study does account for empirical treatment but not secondary transmission. TB NEAT trial did not use mortality or DALYs averted as empirical treatment may result in less impact on mortality.

Shazzadur Rahman employed data from the PROVE-IT study in Brazil and a discrete event simulation model to assess triage strategies and Xpert among persons with presumptive TB at primary health clinics in Brazil (29). Cost per additional true TB patient diagnosed compared to microscopy for Xpert with no triage was US$ 4242, alternative algorithms looked at cough or clinical scores with Xpert but provided no benefit, a neural network approach or improved hypothetical triage approach as triage before Xpert did improve cost-effectiveness. Adding chest X-ray as a triage tool for HIV negative cases could substantially save costs associated with Xpert without triage, and identify almost as many cases.

Lee et al used a modified CEPAC-I model (a microsimulation model) in an HIV negative population in India with presumptive TB to compare cost-effectiveness of 4 strategies including sputum smear microscopy (SSM), Xpert, Truenat and Truenat at point-of-care in primary healthcare facilities (24). Using a willingness to pay threshold of US$ 990/year of life saved, compared to SSM, Truenat POC was cost-effective (ICER US$ 210/YLS), as was centralized Truenat (ICER US$ 240/YLS), Xpert was dominated by POC Truenat. More details on this study from the Truenat test perspective are discussed below.

PICO 3 & 4: Among adults and children with signs and symptoms of extrapulmonary TB, seeking care at health care facilities should Xpert MTB/RIF /Ultra used as an initial test for diagnosis of extrapulmonary TB and Rif resistance?

Only one study by Wang et al assessed cost-effectiveness specifically among extrapulmonary samples in evaluating the incremental cost-effectiveness of a second Xpert assay among presumptive TB patients presenting to the national TB referral center in Beijing China and performed stratified analyses on pulmonary and extrapulmonary samples (26). Average cost per TB case diagnosed was US$ 22.82 for the 1st Xpert and US$ 43.51 for the second Xpert test in pulmonary TB samples and US$ 35.02 for the 1st Xpert and US$ 62.54 for the second Xpert among extrapulmonary cases. Compared to just one Xpert the incremental cost (per incremental case) of performing a second Xpert was US$ 467.72 for pulmonary TB and $291.87 for extrapulmonary TB. While the incremental cost of a second Xpert was substantial, it was beneficial in detecting additional smear negative and RIF- resistant cases. No information was provided regarding whether samples from children were included in this analysis.

PICO 5: Among people with signs and symptoms of pulmonary TB, seeking care at health care facilities does repeated Xpert tests on subsequent samples provide any additional value as an initial test for diagnosis of pulmonary TB and Rif resistance?

Cost-effectiveness of repeated Xpert MTB/RIF Among People living with HIV

Two studies were identified among PLHIV evaluating the use of repeated Xpert tests, both demonstrating additional value of the second Xpert, although Schnippel did demonstrate a slight decrease in number of TB cases diagnosed compared with culture.

As discussed above in reference to PICO 1 & 2, Andrews et al assessed PLHIV initiating ART in South Africa, and compared several diagnostic algorithms both in patients with signs and symptoms of TB and among all patients regardless of symptoms (5). Employing 2 Xperts for all (regardless of symptoms) was more effective with a lower ICER (US$ 5100/YLS) and dominated the single Xpert regardless of symptoms strategy. Authors concluded that performing two Xperts had additional benefit compared with 1 Xpert, and was cost-effective. Costs of Xpert tests represented only 1.2% of care costs in 1 year post TB diagnosis, while the majority of costs are driven by TB and HIV treatment.

Schnippel et al 2013 assessed the addition of a second Xpert test to replace culture among PLHIV with initial negative Xpert tests in South Africa (14). 2 Xperts would result in 2% fewer TB cases and 2% fewer MDR-TB cases being diagnosed due to decreased sensitivity of Xpert compared with culture but these effects were mitigated by 1% more TB cases and 1% more MDR-TB cases initiating treatment due to decreases in loss to follow-up. Furthermore, authors found a decreasing cost per patient treated using 2 Xperts compared with Xpert followed by culture (%6076 v R6435) resulting in an annual cost savings of US$ 17.4 million or 115 of total diagnostic cost.

Cost-effectiveness of repeated Xpert among hospitalized populations

Only one study assessing repeated Xpert was performed in hospitalized patients. As discussed above Cowan et al assessed the use of Xpert amongst hospitalized patients suspected of pulmonary TB in Seattle, USA (21). Both the 1 Xpert and 2 Xpert strategy dominated over 3 SSM, with cost-savings generated from reduction of airborne infection isolation costs in hospital, but 2 the Xpert approach resulted in a very high ICER of US$ 2826682/accurate case diagnosed compared with 1 Xpert.

Cost-effectiveness of repeated Xpert among outpatients

As discussed above Wang et al assessed incremental cost-effectiveness of a second Xpert assay among presumptive TB patients presenting to the national TB referral center in Beijing China (21). Average cost per TB case diagnosed was US$ 22.82 for the 1st Xpert and US$ 43.51 for the second Xpert test. Compared to just one Xpert the incremental cost (per incremental case) of performing a second Xpert was US$ 467.72 for pulmonary TB and US$ 291.87 for extrapulmonary TB. While the incremental cost of a second Xpert was substantial, it was beneficial in detecting additional smear negative and RIF- resistant cases.

PICO 6: Among adults in a population-based TB disease prevalence survey with symptoms or chest X-ray abnormalities suggestive of pulmonary TB, should Xpert MTB/RIF/Ultra alone, be used to define the case of active TB disease?

No studies were identified that addressed cost-effectiveness of Xpert in population-based T disease prevalence survey.

PICO 7: Among people being screened for pulmonary TB, should Xpert MTB/RIF/ Ultra be used alone to define TB and Rif resistance?

Cost-effectiveness of Xpert MTB/RIF for screening among people living with HIV

Two studies were identified addressing screening among PLHIV, both found Xpert to be the most cost-effective approach, although the Andrews analysis found 2 Xperts to be preferred over a single Xpert approach.

Orlando et al assessed screening among ART naïve PLHIV in Mozambique comparing a standard 4 symptom screen followed by SSM with Xpert for all and a third strategy with Urine LF-LAM and Xpert (12). Authors found the Xpert for all approach was most cost-effective with an ICER of US$ 56.54/DALY averted. It should be noted this analysis included cost savings gained through reduction of newly transmitted infections due to delayed diagnosis.

As discussed above in reference to PICO 1, 2, 3 & 4, Andrews et al assessed routine TB screening with Xpert among PLHIV initiating ART in South Africa, and compared several diagnostic algorithms both in patients with signs and symptoms of TB and among all patients regardless of symptoms (5). Employing 2 Xperts for all (regardless of symptoms) was more effective with a lower ICER (US$ 5100/YLS) and dominated strategies including a symptom screen. Authors concluded strategies with symptom screening were less efficient and less cost-effective.

Cost-effectiveness of Xpert for screening amongst high risk groups

Two studies assessed cost-effectiveness of screening with Xpert among high risk groups: one in prisons (30), and the other among elderly persons being admitted to residential care homes (27).

Winetsky et al evaluated cost-effectiveness of 8 different TB screening strategies among prisoners in former Soviet Union countries (30). Using a single Xpert as an annual screening strategy among the general inmate population was the most effective approach in reducing TB and MDR-TB prevalence and resulted in an ICER US$ 543/QALY gained compared to current approach of mass miniature radiography (MMR). Symptom screen strategies alone were less effective and more expensive than current standard of care (MMR). In sensitivity analyses, model results were robust across all parameters evaluated and and Xpert remained cost-effective (using FSU countries average GDP: US$ 10,561 as WTP threshold).

Li et al assessed cost-effectiveness of screening with Xpert among elderly persons being admitted to residential care homes in China (Hong Kong SAR)(27). Compared with passive screening, the Xpert screening approach resulted in an ICER of US$ 6094/QALY gained or US$ 9076/life years saved. Authors found LTBI screening was more cost-effective than TB screening with Xpert when the probability of annual LTBI reactivation was greater than 0.155% and when screening acceptability was greater than 38%.

PICO 8: Among people with signs and symptoms of pulmonary TB, seeking care at health care facilities what is the economic evidence and cost-effectiveness of Molbio TrueNat MTB / Rif to be used as an initial test for diagnosis of pulmonary TB and RR?

Truenat (Molbio Diagnostics/Bigtec Labs, Goa/Bengaluru, India) is a novel molecular assay that offers rapid detection of tuberculosis (TB) and rifampicin-resistance using a battery powered platform which may be potentially useful in peripheral healthcare settings or for point-of-care testing. Only one cost-effectiveness analysis is currently published. Lee et al used a microsimulation model (modified CEPAC-I) to assess cost-effectiveness among an HIV negative population with presumptive TB (cough> 2weeks) and compared 4 strategies including sputum smear microscopy (SSM), Xpert, Truenat and Truenat at point-of-care in primary healthcare facilities (24). Sensitivity of Truenat in the model was 86% slightly below that of Xpert (89%). Truenat unit costs were estimated to be slightly higher than Xpert at US$ 13.20 for Truenat and US$ 12.63 for Xpert. Linkage to care was assumed to be 84% at designated microscopy centers where the first three diagnostic algorithms were modelled and 95% in the POC Truenat algorithm. Using a willingness to pay threshold of US$ 990/year of life saved (USD) representing 50% of the 2017 Indian GDP.

Compared to SSM, Truenat POC increased life-expectancy due to improved linkage to care and treatment initiation and was also cost-effective at an ICER of US$ 210/YLS. Compared to Xpert, Truenat POC increased life-expectancy and was also cost-effective at an ICER of US$ 120/YLS. Key variables included Truenat sensitivity and linkage to care, Truenat's specificity for RIF resistance was most influential, a 10% reduction in Specificity resulted in an increased ICER of US$ 350/YLS. Truenat remained cost-effective even if test volumes decreased by 5 or 10 fold in peripheral clinics. As Truenat's sensitivity decreases the linkage to care necessary to ensure Truenat is cost-effective increases, therefore even suboptimal test employed at peripheral centers may be cost-effective if linkage to care is improved. Neither Xpert nor Truenat POC was cost-effective compared to SSM until 6 years after initial testing was implemented.

Lee et al also performed a budget impact analysis (24). Scaling up Xpert increased TB related healthcare expenditures by US$ 580 million (81% increase) over 2 years, mostly driven by increased

MDR-TB treatment spending. Deploying Truenat POC increased expenditures by an additional $100 million over Xpert (7% increase) over 2 years.

Other economic studies assessing decentralized of testing have suggested decentralized testing approaches (either with Xpert or novel diagnostics) may be cost-effective or even cost-savings but depend largely on testing volume, specimen transport systems, TB prevalence and pre-treatment loss to follow-up (31).

PRINCIPAL FINDINGS

- Studies employed a variety of different modelling approaches, populations and settings. Variations in costing, effectiveness and epidemiological parameters were present across included studies making direct comparisons across studies challenging.

- Studies used both short-term diagnostic outcomes (additional cases diagnosed, RR-TB cases diagnosed) and long-term outcomes (years of life saved, DALYs averted etc.) There was variation in costing elements included across different analyses, both in terms of what was included in unit test costs (consumables and equipment only versus overhead, staffing, training etc.), whether implementation costs were included for introducing novel diagnostic testing, and whether downstream costs associated with TB treatment, MDR-TB treatment and ART and HIV care were included.

- While many studies demonstrated that Xpert may be cost-effective in diagnosing pulmonary TB, key implementation conditions and settings could be largely influential on determining cost-effectiveness and must be considered when implementing Xpert. Cost-effectiveness of Xpert was shown to be improved among populations with higher TB prevalence, among PLHIV populations and where rates of empirical treatment were low.

- Cost-effectiveness of Xpert is highly dependant on a number of important factors including placement of Xpert machines (centralized facilities versus decentralization) test volume, underlying TB prevalence, level of empirical treatment and pre-treatment loss to follow-up.

- No study directly assessing cost-effectiveness of Xpert Ultra were identified. Only one study assessing cost-effectiveness of Molbio's Truenat MTB/RIF was identified, while this study suggests Truenat is likely cost-effective if implemented at the POC in India, it relies on several important modelling assumptions including improved linkage to care and increased treatment initiation.

DISCUSSION

Through a systematic review of the published literature we were able to identify 28 economic studies evaluating Xpert and meeting our inclusion criteria, including one study that also evaluated the cost-effectiveness of Molbio Truenat MTB/RIF in India. Studies were primarily assessing Xpert in African outpatient settings, but also among outpatients in India and Brazil. Four included studies were conducted among hospitalized patients in Germany, China (Hong Kong SAR) and the Unites States of America, 2 screening studies focused on PLHIV in Mozambique and South Africa and one among the elderly in China (Hong Kong SAR) and among prisoners in former Soviet Union countries. One study specifically assessing extrapulmonary TB was conducted in Beijing, China, and repeated

Xperts were evaluated among PLHIV in two studies from South Africa, one among hospitalized patients from the US and one study among outpatients in China. No studies directly assessed the cost-effectiveness of Xpert Ultra.

Included studies highlighted the importance of cost considerations, and how factors around testing volume can have important impacts on cost-effectiveness. Xpert costing studies performed in Uganda, and South Africa have demonstrated higher costs associated with implementing Xpert in rural or peripheral settings due in part to decreased testing volume or increased infrastructure needs in these settings (32,33). Further costing studies including one performed by Naidoo et al using laboratory records and empirical costing in South Africa have demonstrated total TB diagnostic costs increased by 43% from $440 967 during the smear/culture based approach (April- June 2011) to US$ 632 262 using Xpert (April-June 2013) increasing cost per TB case diagnosed by 157% (34). Authors concluded that Xpert resulted in substantial cost increases that were not matched by expected increase in TB efficacy. Studies from Vassall et al using pragmatic trial data have reinforced this notion that expected increases in diagnostic efficacy, case notifications and mortality have not been borne out post- Xpert implementation, although Vassal et al also found implementation of Xpert to be cost-neutral and less expensive compared to initial estimations (16). A study from Pantoja et al also reported using Xpert to diagnose MDR-TB and TB in PLHIV would cost less than conventional diagnostics globally and in all high burden countries, while testing everyone with signs and symptoms with TB would costs much more than conventional diagnostic approaches (35).

While we did not identify any studies assessing cost-effectiveness of Xpert in a prevalence survey, there is costing evidence from a prevalence survey conducted by Dorman et al in South Africa indicating that in a testing scenario of 7000 specimens, total costs for Xpert were $165,690 and US$ 115,360 for microscopy plus culture approach (36). This prevalence study concluded the diagnostic yield with Xpert was substantially higher compared with microscopy but lower than liquid culture in the context of a prevalence survey design, therefore Xpert may be considered for prevalence surveys in settings where liquid culture is not available.

CONCLUSION

While there is a substantial amount of economic evidence around implementation and scale-up of Xpert in a variety of settings, most notably among outpatients presenting with signs and symptoms of TB, and the majority of studies found Xpert could be likely cost-effective, not all were consistent in this finding and studies highlighted differences in implementation approaches and settings could have important impact on cost-effectiveness results. Cost-effectiveness of Xpert was shown to be improved among populations with higher TB prevalence, among PLHIV and where rates of empirical treatment were low, and linkage to care or pre-treatment loss to follow-up were poor. Studies employed a wide variety of modelling and analysis approaches, assumptions, diagnostic algorithms, and comparators, and assessed different study settings making comparisons across studies and generalizations to other settings challenging.

Studies highlighted implementation factors and setting should be considered an important element when generalizing cost-effectiveness results to different settings. Considerations regarding current standard of care, level of empirical treatment, existing testing facilities, placement of Xpert (peripheral or point of care settings versus centralized facilities) TB prevalence, patient volume, pre-treatment loss to follow-up and existing linkage to care are all important factors in determining whether Xpert may be cost-effective in any given setting. Studies also highlighted importance of cost components including whether implementation costs associated with Xpert scale-up were considered and whether downstream costs such as TB and MDR-TB treatment and ART and HIV care costs were included.

Economic evidence regarding the implementation and scale-up of Molbio's Truenat MTB/RIF is very limited with only one published study available, while this study suggests Truenat is likely cost-effective if implemented at the POC in India, it relies on several important modelling assumptions including improved linkage to care and increased treatment initiation which should be evaluated in pragmatic trials as has been done for Xpert implementation in South Africa.

In conclusion, caution should be used when generalizing cost-effectiveness and economic evaluations across settings. Local implementation conditions and settings should be considered and local implementation studies may be helpful to assess likely impact on case-finding, long-term outcomes and cost-effectiveness.

Table 1. Characteristics of included studies

Study Characteristics	Abimbola 2012, JAIDS	Adelman 2018, OFID	Andrews 2012, AIDS
Country setting	Sub-Saharan Africa	Ethiopia	South Africa
Year of cost valuation	2010	2014	2010
Currency	USD	USD	USD
Clinical setting	Not Specified	Outpatient HIV clinic	Outpatient
Study population	Patients with advanced HIV initiating ART with signs and symptoms of TB.	PLHIV presenting to HIV clinic	HIV-infected individuals initiating ART
Xpert diagnostic strategies	Xpert MTB/RIF	WHO recommended symptom screening followed by Xpert in positive symptom screen	• Single sputum Xpert • Two concurrent sputum samples tested with Xpert (Looked at testing among symptomatic persons versus testing all)
Reference diagnostic strategies	Sputum smear microscopy (2 samples) with smear negative individuals undergoing CXR.	WHO recommended symptom screening followed by 3 smears in persons with symptoms	2 Cultures for all
Analysis perspective	Health system	Health system	Health systems
Type of economic evaluation	CEA	CUA	CEA
Source of costing	Published literature	Empirical data collection & published literature	Empirical data collection & published literature
Primary outcome	ICER: $/ Deaths averted	ICER: $/ DALYs averted	ICER: $/ year of life saved

Type of model	Decision analytic	Decision analytic	Monte Carlo microsimulation model (Modified CEPAC-I)
Sensitivity analyses	Univariable and probabilistic	Univariable	Univariable and two way
Key scenarios/variables explored in sensitivity analyses	Univariable: all parameters; Probabilistic: disease prevalence parameters, mortality rates and cost inputs	All model inputs and costs: Xpert cartridge cost, MDR treatment cost, high TB prevalence, cost of SSM & Sensitivity of SSM	Univariable: all model input parameters; Two-way: sensitivities of smear and Xpert, spectrum of decreasing test costs over time
WTP threshold	US$ 5,678 (per capita GDP of South Africa)	US$ 505 x3 (3x per capita GDP of Ethiopia)	US$ 7,100 x3 (3x per capita GDP of South Africa)

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	Cowan 2017, CID	Diel 2016, Eur Respir J	Dunbar 2018, IJTLD
Country setting	USA	Germany	South Africa
Year of cost valuation	2015	2013	2013
Currency	USD	Euros (€)	USD
Clinical setting	Inpatient	Inpatient	Outpatient
Study population	Admitted patients evaluated for PTB	Untreated TB suspects	Presumptive TB cases presenting to primary health clinics
Xpert diagnostic strategies	• 1 Xpert on an unconcentrated sputum sample; • 1 Xpert on a concentrated sputum sample; • 2 consecutive Xperts on a concentrated sputum sample	Xpert MTB/RIF on a single sputum sample, followed by a culture **Note:** Other Xpert add-on strategies addressed, but do not meet our selection criteria	Xpert-based algorithm: Xpert as first test followed by smear/culture/LPA as needed for additional drug sensitivity testng/ or according to HIV status
Reference diagnostic strategies	• 2 consecutive smears • 3 consecutive smears	Sputum smear and culture	Smear-based algorithm: smear as first test and additional smear or culture as needed or according to HIV status
Analysis perspective	Hospital perspective	Hospital perspective	Health system
Type of economic evaluation	CEA	CBA	CEA
Source of costing	Empirically collected and published literature	Published literature	Published literature
Primary outcome	ICER: $/accurately diagnosed case	Mean incremental cost/patient	$/RR-TB case identified $/additional RR-TB case identified
Type of model	Decision analytic	Decision analytic	Discrete event simulation/operational model
Sensitivity analyses	Univariable, two-way, and probabilistic	Univariable and probabilistic	Univariable
Key scenarios/variables explored in sensitivity analyses	All variables	TB Prevalence & MDR Prevalence, Costs of hospitalization and isolation	• Varied levels of adherence to the Xpert algorithm (at increments of 10%, from 50% to 100%) • Varied the proportion of presumptive TB cases who knew their HIV status (at 60%, 80% and 100%),

WTP threshold	US$ 50,000	Not reported	Not reported
Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CBA, cost-benefit analysis; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved			

Study Characteristics	Dunbar 2017, IJTLD	Jha 2016, PloS One	Khaparde 2017, PloS One
Country setting	South Africa	South Africa	India
Year of cost valuation	2013	2015	2013
Currency	USD	USD	USD
Clinical setting	Outpatient	Outpatient	Outpatient
Study population	Presumptive TB cases	Adults with clinical suspicion of TB	Adult patients with signs and symptoms of pulmonary TB
Xpert diagnostic strategies	Xpert-based algorithm: Xpert as first test followed by smear/culture/LPA as needed for additional drug sensitivity testng/ or according to HIV status	Xpert MTB/RIF performed on all specimens	Upfront Xpert MTB/RIF for all presumptive TB patients **Note**. Other Xpert add-on strategies addressed, but do not meet our selection criteria
Reference diagnostic strategies	Smear-based algorithm: smear as first test and additional smear or culture as needed or according to HIV status	Sputum smear microscopy alone	• Sputum smear microscopy for all presumptive TB patients; • Xpert MTB/ RIF for presumptive TB cases with previous TB history, sputum smear microscopy for new patients.
Analysis perspective	Health system	Health system	Health system
Type of economic evaluation	CEA	CEA	CEA
Source of costing	Previous costing evaluation and published literature	Published literature	Observational micro-costing study and published literature
Primary outcome	• $/RR-TB case identified • $/additional RR-TB case identified	ICER: $/ true-positive diagnosis made	• $/presumptive TB patient tested • $/true TB case detected and initiated on treatments
Type of model	Discrete event simulation/Operational model	Decision analytic model	Decision analytic model
Sensitivity analyses	Scenario analysis	Univariable & probabilistic sensitivity analysis	Deterministic sensitivity analyses

Key scenarios/variables explored in sensitivity analyses	• Varied TB prevalence among presumptive cases being tested • Cost per TB case diagnosed if the price per Xpert cartridge was reduced by 10%, 25% and 50%	• One-way and probabilistic: all model parameters • Scenarios: high- and low-volume setting and at different assumed levels of MDR-TB prevalence	Epidemiological and cost parameters
WTP threshold	Not reported	US$ 1927/incremental diagnosis	Not reported
Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved			

Study Characteristics	**Langley 2014, Lancet**	**Lee 2019, PloS One**	**Li 2018, PloS One**
Country setting	Tanzania	India	China (Hong Kong SAR)
Year of cost valuation	2012	2017	Not stated
Currency	USD	USD	USD
Clinical setting	Outpatient	Outpatient	Outpatient
Study population	Patients with presumptive TB	HIV-negative adults with presumptive TB (cough of at least 2 weeks duration)	65-year-old elderly population at admission to residential care homes for the elderly
Xpert diagnostic strategies	• Xpert followed by clinical judgement (including chest x-ray) if smear negative, for • For all presumptive TB cases, or • Known HIV-positive patients	• Xpert in designated microscopy centers; • Truenat in designated microscopy centers; • Truenat for point-of-care testing in primary healthcare facilities	Xpert on patients with symptoms of TB at admission to residential care homes
Reference diagnostic strategies	Sputum smear microscopy followed by clinical judgement if smear negative	Sputum smear microscopy in designated microscopy centers (centralized facilities)	No screening
Analysis perspective	Health system	Health system	Health system

Type of economic evaluation	CUA	CEA	CUA
Source of costing	Published literature	Empirical data collection & published data	Published literature & public data
Primary outcome	ICER: US$ /DALYs averted	ICER: US$ /YLS	ICER: US$ /QALYs gained
Type of model	Operational model	Microsimulation model	Decision analytic Markov model
Sensitivity analyses	Uncertainty analysis and univariable sensitivity analysis	Univariable, two-way, and scenario analyses	Univariable and probabilistic
Key scenarios/variables explored in sensitivity analyses	Uncertainty: population-level effect on incidence, prevalence, mortality, and DALYs; One-way: model input variables	One-way: model parameters in Table 1; Two-way: Truenat's sensitivity for TB and linkage-to-care at 5-year horizon; Scenario analyses: effect of empirical treatment on cost-effectiveness, differential loss-to-follow up, and per-test cost of Truenat	Key model input parameters
WTP threshold	US$ 599 (per-capita GDP of Tanzania in 2012)	US$ 990/YLS (50% of India's per-capita GDP in 2017)	US$ 50,000

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	**Menzies 2012, PloS Medicine**	**Meyer-Rath 2012, PloS One**	**Millman 2013, PloS One**
Country setting	Botswana, Lesotho, Namibia, South Africa, and Swaziland	South Africa	United States
Year of cost valuation	2011	2011	2009
Currency	USD	USD	USD
Clinical setting	Outpatient	Outpatient & inpatient	Inpatient
Study population	Patients with presumptive TB	Adult patients with suspected pulmonary TB	Inpatients who underwent microbiologic testing for TB while in respiratory isolation
Xpert diagnostic strategies	Xpert as an initial diagnostic test for all patients with suspected TB	• Xpert as initial test, scaled up by the end of 2012; • Xpert as initial test, scaled up	Xpert testing of a single sputum sample

		by the end of 2013. Xpert followed by smear, culture, line-probe assay, and drug susceptibility testing depending on HIV status in all strategies	
Reference diagnostic strategies	Sputum smear culture with smear-positive directed to treatment and smear-negative undergoing culture if there is a history of TB-treatment or strong suspicion of TB	Sputum smear microscopy followed by culture, line probe assay, and drug susceptibility testing based on HIV status	Two concurrent SSM
Analysis perspective	Health system	Health system	Health system
Type of economic evaluation	CUA	CEA	CEA
Source of costing	Published literature	Expert opinion, public-sector salary data & published literature	Empirical data collection & published literature
Primary outcome	ICER: $/DALYs averted	• $/ case diagnosed and treated • incremental cost per case	Incremental net monetary benefit of the Xpert strategy relative to the smear strategy
Type of model	Dynamic compartmental model	Population-level decision model	Decision analytic
Sensitivity analyses	Univariable, probabilistic, Bayesian uncertainty analysis, additional sensitivity analyses	Univariable	Univariable, two-way, and probabilistic
Key scenarios/variables explored in sensitivity analyses	One-way and Bayesian: model input parameters; Probabilistic: choice of diagnostic strategy from the joint effects of uncertainty around all input parameters; Additional: varied assumptions regarding the diagnostic algorithms, inpatient care as part of MDR-TB treatment, ART coverage and drug prices.	• Impact of full Xpert coverage on smear positivity and culture positivity rates of suspects as a result of a reduction in transmission; • Xpert cartridges at the volume-discounted price;	Epidemiological and cost input parameter

		• additional 4 months of inpatient care per patient for MDR-TB	
WTP threshold	US$ 982 to US$ 7,000 (per-capita GDP in each	Not stated	Not stated

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	Orlando 2018, PloS One	Pooran 2019, Lancet Glob Health	Schnippel 2013, SAMJ
Country setting	Mozambique	South Africa, Zambia, Zimbabwe and Tanzania	South Africa
Year of cost valuation	2016	2014	2011
Currency	USD	USD	ZAR (South-African Rand)
Clinical setting	Inpatients & outpatients	Outpatient	Outpatient
Study population	PLHIV	Individuals with presumptive TB	Individuals with presumptive TB
Xpert diagnostic strategies	Xpert MTB/RIF for all patients	Same-day Xpert MTB/RIF at clinic (POC) **Note**. An additional Xpert was performed on a stored sputum sample at a centralised laboratory (Lab Xpert) by a qualified technician- not evaluated in cost-effectiveness.	Initial Xpert followed by a second Xpert if the first is negative
Reference diagnostic strategies	Four symptom screen with smear for participants with positive screen results	Two sputum samples tested with smear microscopy followed by liquid culture	Culture
Analysis perspective	Health system	Health system	Health system
Type of economic evaluation	CUA	CEA	CEA
Source of costing	DREAM program & published literature	Empirically collected at each individual trial site	Xpert implementation studies and public sector price and salary data
Primary outcome	ICER: $/DALYs averted	• Incremental cost/treatment initiation • Incremental cost/treatment completion	ICER: ZAR/ case diagnosed)
Type of model	Decision analytic	Not a modelling study	Population-level decision model
Sensitivity analyses	Univariable	Univariable & probabilistic	Univariable
Key scenarios/variables explored in sensitivity analyses	Key model input parameters	Univariable: model input parameters; Probabilistic: simultaneous varying of cost and effectiveness parameter inputs Scenarios: incremental cost per treatment initiation and the incremental cost per	Systematically varied eight central parameters: Xpert sensitivity for smear-negative TB, cost of the Xpert, proportion of patients with possible TB who have known HIV infection; proportion of patients

		treatment completion among culture positive patients	lost at each visit, proportion of TB which is smear positive, TB positivity rate, access to LPA testing, proportion testing rifampicin resistant
WTP threshold	US$ 1,146 (per-capita GDP of Mozambique)	Not stated	Not stated

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	ShazzadurRahman 2019, BMC ID	Suen 2015, IJTLD	Tesfaye 2017, BMC ID
Country setting	Brazil	India	Ethiopia
Year of cost valuation	Not stated	2013	2015
Currency	USD	USD	USD
Clinical setting	Outpatient	Outpatient	Outpatient
Study population	Patients with symptoms or signs suggestive of TB	Individuals with presumptive TB	Patients with presumptive TB
Xpert diagnostic strategies	(1) Xpert with no triage; (2) Xpert with >1 week of cough as triage; (3) Xpert with >3 weeks of cough as triage	Xpert for initial TB diagnosis and DST Note. Other Xpert add-on strategies addressed, but do not meet our selection criteria	All patients with presumptive TB tested with Xpert Note. Other Xpert add-on strategies addressed, but do not meet our selection criteria
Reference diagnostic strategies	SSM based on two samples collected on different days followed by a clinical assessment for smear negative cases	SSM	3 concurrent SSM
Analysis perspective	Health system	Societal	Health system
Type of economic evaluation	CEA	CUA	CUA
Source of costing	Published literature	Published literature	Empirical data collection
Primary outcome	$/ additional true TB patient diagnosed and treated	ICER: $/QALYs gained	ICER: $/DALYs averted
Type of model	Operational model	Dynamic transmission microsimulation model	Operational/ discrete event simulation
Sensitivity analyses	Univariable	Univariable, multivariable, probabilistic, and scenario analyses	Probabilistic
Key scenarios/variables explored in sensitivity analyses	Accuracy (sensitivity and specificity) of the triage tool, lower prevalence of active TB, cost of triage	Xpert and public-private mix attributes & simultaneous effect of uncertainty on the quality of life lost due to TB and the costs of care	Model input parameters
WTP threshold	Not stated	US$ 1,450 (per-capita GDP of India)	US$ 690 (per-capita GDP of Ethiopia)

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; DST, drug sensitivity testing; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	Vassall 2017, Lancet Glob Health	Vassall 2011, PLoS medicine	Wang 2018, J Thorac Dis
Country setting	South Africa	India, Uganda, South Africa	China
Year of cost valuation	2014	2010	Not stated
Currency	USD	USD	USD
Clinical setting	Outpatient	Outpatient	Not stated
Study population	People being assessed for tuberculosis attending primary health-care clinics	Individuals with presumptive TB	Pulmonary TB suspects and extrapulmonary TB suspects who had two Xpert tests sequentially within one week
Xpert diagnostic strategies	Sputum Xpert as an initial test for TB	Single sputum specimen tested by Xpert for all individuals Note. Other Xpert add-on strategies addressed, but do not meet our selection criteria	• Single Xpert • Two concurrent Xperts
Reference diagnostic strategies	SSM	Two SSM, followed by CXR in smear negative	SSM
Analysis perspective	Societal	Health system	Health system
Type of economic evaluation	CUA	CUA	CEA
Source of costing	Empirical data collection from XTEND trial	Empirical data collection	Empirical data collection
Primary outcome	ICER: $/DALYs averted	ICER: $/DALY averted	$/ case detected Incremental $/ additional TB case identified
Type of model	Not a modelling study	Decision analytic model	Decision analytic
Sensitivity analyses	One-way sensitivity analysis	One-way, two-way, and probabilistic	None
Key scenarios/variables explored in sensitivity analyses	Discount rates	Epidemiological and cost input parameters	N/A
WTP threshold	US$ 0- US$ 10,000	Per capita GDP by country: India: US$ 1,134 Uganda: US$ 490; South Africa: $5,786	Not stated

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; DST, drug sensitivity testing; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	Wikman-Jorgensen 2017, TROP MED INT HEALTH	Winetsky 2012, PLoS medicine	You 2015, J. Infect
Country setting	Mozambique	Former Soviet Union	China (Hong Kong SAR)
Year of cost valuation	2013	2009	2014
Currency	USD	USD	USD
Clinical setting	Outpatient	General Prison Population	Inpatient
Study population	Persons with presumptive TB	Inmates in prisons of the Former Soviet Union	Adult patients hospitalized for suspected active pulmonary TB
Xpert diagnostic strategies	Xpert	Xpert	Xpert
Reference diagnostic strategies	Two SSM followed by a chest X-ray or antibiotic trial in smear-negative TB suspects	MMR screening with sputum PCR detection of MDR-TB	• Two sputum microscopy examinations, with smear-negative patients receiving clinical diagnosis • Two sputum microscopy examinations, with smear-negative patients tested by Xpert
Analysis perspective	Health system	Health system	Health system
Type of economic evaluation	CUA	CUA	CUA
Source of costing	Empirical data collection & published literature	Empirical data collection from Tajikistan and published literature	China (Hong Kong SAR) Hospital Authority
Primary outcome	ICER: $/DALYs averted	ICER: $/QALYs gained	ICER: $/ QALYs gained
Type of model	Stochastic Markov model	Deterministic, population-based compartmental model	Decision analytic model
Sensitivity analyses	Univariable & probabilistic	Univariable, two-way, & probabilistic	Univariable & probabilistic
Key scenarios/variables explored in sensitivity analyses	All model parameters	Model parameter estimates	All model parameters

WTP threshold	$590 (per-capita GNI of Mozambique)	Per-capita GDP of the Former Soviet Union	$50,000

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; DST, drug sensitivity testing; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Study Characteristics	**Zwerling 2015, JAIDS**
Country setting	Malawi
Year of cost valuation	2010
Currency	USD
Clinical setting	Outpatient
Study population	People newly diagnosed with HIV
Xpert diagnostic strategies	4 symptom screen followed by Xpert in those with any symptom
Reference diagnostic strategies	4 symptom screen followed by clinical judgement of the treating physician
Analysis perspective	Health system
Type of economic evaluation	CUA
Source of costing	Empirical data collection
Primary outcome	ICER: $/DALYs averted
Type of model	Decision analytic
Sensitivity analyses	Univariable, two-way, probabilistic, and scenario analyses
Key scenarios/variables explored in sensitivity analyses	cost-effectiveness under conditions of high, medium, and low test volume, with and without ART, and across varying levels of symptom-driven diagnosis of TB in the standard of care
WTP threshold	$1417 (average per capita GDP of low-income countries in Sub-Saharan Africa)

Abbreviations: ART, antiretroviral therapy; CXR, chest x-ray; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; DALY, disability adjusted life year; DST, drug sensitivity testing; GDP, gross domestic product; GNI, gross national income; IPT, Isoniazid preventive therapy; PLHIV, people living with HIV; TB, tuberculosis; USD, United States dollars; Xpert, Xpert MTB/RIF; YLS, years of life saved

Table 2. Cost components for unit test cost estimation

		Xpert costs									Treatment costs
First Author, Year, Journal	Country	Lab space	Staff	Training	Equipment	Consumable	Overhead	Disposal	Transport	Cost of Xpert test[1]	Cost of TB treatment[1]
Abimbola 2012, JAIDS	Sub-Saharan					✓				US$ 31.65	
Adelman 2018, OFID	Ethiopia				✓	✓					US$ 33 (US$ 4856 for MDR-
Andrews 2012, AIDS	South Africa					✓				US$ 21.60	US$ 6.60-$140.00
Cowan 2017, CID	United States		✓		✓	✓	✓			US$ 116.00	US$ 50.21/da
Diel 2016, Eur Respir J	Germany					✓				€110.75	€6.3/day (€101.04/day for MDR-TB)
Dunbar 2018, INT J TUBERC LUNG DIS	South Africa	✓	✓		✓	✓	✓			US$ 19.03	
Dunbar 2017, INT J TUBERC LUNG DIS	South Africa	✓	✓		✓	✓	✓			US$ 19.03	
Jha 2016, PloS One	South Africa	✓	✓	✓	✓	✓	✓			US$ 14.45 - US$ 16.6	US$ 506 ($3660 for MDR-
Khaparde 2017, PloS One	India					✓				US$ 13.17	US$ 28.13 - US$ 104.23
Langley 2014, Lancet	Tanzania		✓		✓	✓					US$ 3.00 - US$ 119.40 per month
Lee 2019, PloS One	India	✓	✓		✓	✓	✓			US$ 12.63 (Xpert) US$ 13.20 (TrueNat)	US$ 28.13 - US$ 104.23
Li 2018, PloS One	China		✓			✓				US$ 128	US$ 162.00 per 6

Menzies 2012, PloS Medicine	Botswana, Lesotho, Namibia, South Africa, and Swaziland	✓	✓		✓	✓	✓		✓	US$ 20.00 - US$ 40.00	US$ 5.86 - US$ 179.06 per month

Abbreviations: ARV, antiretroviral; TB, tuberculosis; Xpert, Xpert MTB/RIF.
[1]Costs in USD unless stated otherwise
'✓' indicate cost component was explicitly included in unit test cost calculation

		Xpert costs									**Treatment costs**
First Author, Year, Journal	**Country**	**Lab space**	**Staff**	**Training**	**Equipment**	**Consumable**	**Overhead**	**Disposal**	**Transport**	**Cost of Xpert test[1]**	**Cost of TB treatment[1]**
Meyer-Rath 2012, PloS One	South Africa		✓	✓		✓	✓	✓	✓	US$ 32.00	US$ 429.00 - US$ 20,530.
Millman 2013, PloS One	United States		✓		✓	✓				US$ 218.	US$ 4.55 per day
Orlando 2018, PloS One	Mozambique		✓		✓	✓				US$ 14.72	US$ 9.84 per
Pooran 2019, Lancet Glob Health	South Africa, Zambia, Zimbabwe, Tanzania		✓		✓	✓	✓		✓	US$ 24.74- US$ 35.70	US$ 2.05 - US$ 8.07
Schnippel 2013, SAMJ	South Africa					✓	✓		✓	R166.20	R2,768.00 - R205,910.00
ShazzadurRahman 2019, BMC Infect. Dis.	Brazil		✓		✓	✓				US$ 17.80	US$ 840
Suen 2015, INT J TUBERC LUNG DIS	India					✓				US$ 18.30	US$ 840 per TB case (US$ 6313 per MDR-
Tesfaye 2017, BMC Infect. Dis.	Ethiopia		✓		✓	✓				US$ 9.98	US$ 3 - US$ 18.80 per month (US$ 199.40 per month
Vassall 2017, Lancet Glob Health	South Africa	✓	✓	✓	✓	✓				US$ 24.42	US$ 171.12 - US$ 252.9

											for MDR-TB)

Abbreviations: ARV, antiretroviral; TB, tuberculosis; Xpert, Xpert MTB/RIF.
[1]Costs in USD unless stated otherwise
'✓' indicate cost component was explicitly included in unit test cost calculation

		Xpert costs									Treatment costs
First Author, Year, Journal	**Country**	**Lab space**	**Staff**	**Training**	**Equipment**	**Consumable**	**Overhead**	**Disposal**	**Transport**	**Cost of Xpert test[1]**	**Cost of TB treatment[1]**
Vassall 2011, PLoS medicine	India, Uganda, South Africa	✓	✓		✓	✓	✓			US$ 22.63 (India), US$ 25.90 (South Africa), US$ 27.55	US$ 227 (India), US$ 454 (South Africa) $185
Wang 2018, J Thorac Dis	China	✓	✓		✓	✓				US$ 13.20	
Wikman-Jorgensen 2017, TROP MED INT HEALTH	Mozambique	✓	✓	✓	✓	✓			✓	US$ 12.92 reagents pr patient only	US$ 68.13 per TB case treated
Winetsky 2012, PLoS medicine	Former Soviet Union (FSU)		✓		✓	✓	✓			US$ 24.08	US$ 364.45 - $7961.02
You 2015, J. Infect	China (Hong Kong SAR)		✓			✓				US$ 128	US$ 27 per month (US$ 769 per month for
Zwerling 2015, JAIDS	Malawi	✓	✓	✓	✓	✓	✓			US$ 90.50 assuming 100 tests/year	US$ 185.00 ($1739.00 for second-

Abbreviations: ARV, antiretroviral; TB, tuberculosis; Xpert, Xpert MTB/RIF.
[1]Costs in USD unless stated otherwise
'✓' indicate cost component was explicitly included in unit test cost calculation

References

1. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. PLOS Med. 2009 Jul 21;6(7):e1000100.

2. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7):e1000097.

3. Abimbola T, Marston B, Date A, Blandford J, Sangrujee N, Wiktor S. Cost-Effectiveness of Tuberculosis Diagnostic Strategies to Reduce Early Mortality Among Persons With Advanced HIV Infection Initiating Antiretroviral Therapy. Jaids J Acquir Immune Defic Syndr. 2012 May 1;60(1).

4. Adelman MW, McFarland DA, Tsegaye M, Aseffa A, Kempker RR, Blumberg HM. Cost-effectiveness of WHO-Recommended Algorithms for TB Case Finding at Ethiopian HIV Clinics. Open Forum Infect Dis. 2018 Jan 1;5(1).

5. Andrews J, Lawn S, Rusu C, Wood R, Noubary F, Bender M, et al. The cost-effectiveness of routine tuberculosis screening with Xpert MTB/RIF prior to initiation of antiretroviral therapy: a model-based analysis. Aids. 2012 May 15;26(8):987–95.

6. Dunbar R, Naidoo P, Beyers N, Langley I. High laboratory cost predicted per tuberculosis case diagnosed with increased case finding without a triage strategy. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2017 01;21(9):1026–34.

7. Dunbar R, Naidoo P, Beyers N, Langley I. Improving rifampicin-resistant tuberculosis diagnosis using Xpert® MTB/RIF: modelling interventions and costs. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2018 01;22(8):890–8.

8. Jha S, Ismail N, Clark D, Lewis JJ, Omar S, Dreyer A, et al. Cost-Effectiveness of Automated Digital Microscopy for Diagnosis of Active Tuberculosis. PLOS ONE. 2016 Jun 20;11(6):e0157554.

9. Langley I, Lin H-H, Egwaga S, Doulla B, Ku C-C, Murray M, et al. Assessment of the patient, health system, and population effects of Xpert MTB/RIF and alternative diagnostics for tuberculosis in Tanzania: an integrated modelling approach. Lancet Glob Health. 2014 Oct 1;2(10):e581–91.

10. Menzies NA, Cohen T, Lin H-H, Murray M, Salomon JA. Population Health Impact and Cost-Effectiveness of Tuberculosis Diagnosis with Xpert MTB/RIF: A Dynamic Simulation and Economic Evaluation. PLOS Med. 2012 Nov 20;9(11):e1001347.

11. Meyer-Rath G, Schnippel K, Long L, MacLeod W, Sanne I, Stevens W, et al. The Impact and Cost of Scaling up GeneXpert MTB/RIF in South Africa. PLOS ONE. 2012 May 31;7(5):e36966.

12. Orlando S, Triulzi I, Ciccacci F, Palla I, Palombi L, Marazzi MC, et al. Delayed diagnosis and treatment of tuberculosis in HIV+ patients in Mozambique: A cost-effectiveness analysis of screening protocols based on four symptom screening, smear microscopy, urine LAM test and Xpert MTB/RIF. PLOS ONE. 2018 Jul 19;13(7):e0200523.

13. Pooran A, Theron G, Zijenah L, Chanda D, Clowes P, Mwenge L, et al. Point of care Xpert MTB/RIF versus smear microscopy for tuberculosis diagnosis in southern African primary care clinics: a multicentre economic evaluation. Lancet Glob Health. 2019 Jun 1;7(6):e798–807.

14. Schnippel K, Meyer-Rath G, Long L, Stevens WS, Sanne I, Rosen S. Diagnosing Xpert MTB/RIF negative TB: Impact and cost of alternative algorithms for South Africa. S Afr Med J. 2013 Jan 14;103(2):101–6.

15. Tesfaye A, Fiseha D, Assefa D, Klinkenberg E, Balanco S, Langley I. Modeling the patient and health system impacts of alternative xpert® MTB/RIF algorithms for the diagnosis of pulmonary tuberculosis in Addis Ababa, Ethiopia. BMC Infect Dis. 2017 May 2;17(1):318.

16. Vassall A, Siapka M, Foster N, Cunnama L, Ramma L, Fielding K, et al. Cost-effectiveness of Xpert MTB/RIF for tuberculosis diagnosis in South Africa: a real-world cost analysis and economic evaluation. Lancet Glob Health. 2017 Jul 1;5(7):e710–9.

17. Wikman-Jorgensen PE, Llenas-García J, Pérez-Porcuna TM, Hobbins M, Ehmer J, Mussa MA, et al. Microscopic observation drug-susceptibility assay vs. Xpert® MTB/RIF for the diagnosis of tuberculosis in a rural African setting: a cost–utility analysis. Trop Med Int Health. 2017;22(6):734–43.

18. Zwerling A, Sahu M, Ngwira L, Khundi M, Harawa T, Corbett E, et al. Screening for Tuberculosis Among Adults Newly Diagnosed With HIV in Sub-Saharan Africa: A Cost-Effectiveness Analysis. Jaids J Acquir Immune Defic Syndr. 2015 Sep 1;70(1):83–90.

19. Vassall A, Kampen S van, Sohn H, Michael JS, John KR, Boon S den, et al. Rapid Diagnosis of Tuberculosis with the Xpert MTB/RIF Assay in High Burden Countries: A Cost-Effectiveness Analysis. PLOS Med. 2011 Nov 8;8(11):e1001120.

20. Diel R, Nienhaus A, Hillemann D, Richter E. Cost–benefit analysis of Xpert MTB/RIF for tuberculosis suspects in German hospitals. Eur Respir J. 2016 Feb;47(2):575–87.

21. Cowan JF, Chandler AS, Kracen E, Park DR, Wallis CK, Liu E, et al. Clinical Impact and Cost-effectiveness of Xpert MTB/RIF Testing in Hospitalized Patients With Presumptive Pulmonary Tuberculosis in the United States. Clin Infect Dis. 2017 Feb 15;64(4):482–9.

22. Millman AJ, Dowdy DW, Miller CR, Brownell R, Metcalfe JZ, Cattamanchi A, et al. Rapid Molecular Testing for TB to Guide Respiratory Isolation in the U.S.: A Cost-Benefit Analysis. PLOS ONE. 2013 Nov 20;8(11):e79669.

23. Suen S-C, Bendavid E, Goldhaber-Fiebert JD. Cost-effectiveness of improvements in diagnosis and treatment accessibility for tuberculosis control in India. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2015 Sep;19(9):1115–24, i–xv.

24. Lee DJ, Kumarasamy N, Resch SC, Sivaramakrishnan GN, Mayer KH, Tripathy S, et al. Rapid, point-of-care diagnosis of tuberculosis with novel Truenat assay: Cost-effectiveness analysis for India's public sector. PLOS ONE. 2019 Jul 2;14(7):e0218890.

25. Khaparde S, Raizada N, Nair SA, Denkinger C, Sachdeva KS, Paramasivan CN, et al. Scaling-up the Xpert MTB/RIF assay for the detection of tuberculosis and rifampicin resistance in India: An economic analysis. PLOS ONE. 2017 Sep 7;12(9):e0184270.

26. Wang G, Wang S, Jiang G, Fu Y, Shang Y, Huang H. Incremental cost-effectiveness of the second Xpert MTB/RIF assay to detect Mycobacterium tuberculosis. J Thorac Dis. 2018 May 3;10(3):1689–95.

27. Li J, Yip BHK, Leung C, Chung W, Kwok KO, Chan EYY, et al. Screening for latent and active tuberculosis infection in the elderly at admission to residential care homes: A cost-effectiveness analysis in an intermediate disease burden area. PLOS ONE. 2018 Jan 2;13(1):e0189531.

28. You JHS, Lui G, Kam KM, Lee NLS. Cost-effectiveness analysis of the Xpert MTB/RIF assay for rapid diagnosis of suspected tuberculosis in an intermediate burden area. J Infect. 2015 Apr 1;70(4):409–14.

29. Shazzadur Rahman AAM, Langley I, Galliez R, Kritski A, Tomeny E, Squire SB. Modelling the impact of chest X-ray and alternative triage approaches prior to seeking a tuberculosis diagnosis. BMC Infect Dis. 2019 Jan 28;19(1):93.

30. Winetsky DE, Negoescu DM, DeMarchis EH, Almukhamedova O, Dooronbekova A, Pulatov D, et al. Screening and Rapid Molecular Diagnosis of Tuberculosis in Prisons in Russia and Eastern Europe: A Cost-Effectiveness Analysis. PLOS Med. 2012 Nov 27;9(11):e1001348.

31. Sohn H, Kasaie P, Kendall E, Gomez GB, Vassall A, Pai M, et al. Informing decision-making for universal access to quality tuberculosis diagnosis in India: an economic-epidemiological model. BMC Med. 2019 Aug 6;17(1):155.

32. Hsiang E, Little KM, Haguma P, Hanrahan C, Katamba A, Cattamanchi A, et al. Higher cost of implementing Xpert® MTB/RIF in Ugandan peripheral settings: Implications for cost-effectiveness. Int J Tuberc Lung Dis. 2016 Sep 1;20(9).

33. Schnippel K, Meyer-Rath G, Long L, MacLeod W, Sanne I, Stevens W, et al. Scaling up Xpert MTB/RIF technology: the costs of laboratory- vs.clinic-based roll-out in South Africa. Trop Med Int Health. 2012 Sep;17(9):1142–51.

34. Naidoo P, Dunbar R, du Toit E, van Niekerk M, Squire SB, Beyers N, et al. Comparing laboratory costs of smear/culture and Xpert® MTB/RIF-based tuberculosis diagnostic algorithms. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2016;20(10):1377–85.

35. Pantoja A, Fitzpatrick C, Vassall A, Weyer K, Floyd K. Xpert MTB/RIF for diagnosis of tuberculosis and drug-resistant tuberculosis: a cost and affordability analysis. Eur Respir J. 2013 Sep 1;42(3):708–20.

36. Dorman SE, Chihota VN, Lewis JJ, Shah M, Clark D, Grant AD, et al. Performance Characteristics of the Cepheid Xpert MTB/RIF Test in a Tuberculosis Prevalence Survey. PLOS ONE. 2012 Aug 15;7(8):e43307.

Web Annex D.6. Report on user perspectives on Xpert testing: results from qualitative research

Authors: Nora Engel & Muthoni Mwaura
Department of Health, Ethics & Society; Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands (Kingdom of the)
n.engel@maastrichtuniversity.nl

Produced in preparation for the WHO guideline group meeting on "**Molecular assays intended as initial tests for the diagnosis of pulmonary and extrapulmonary TB in adults and children. Policy Update",**
3-6 December 2019.

1. Introduction

In ensuring access to effective diagnostics for TB care, we not only need to assess that these technologies are accurate but also that they are feasible, useable and acceptable. The users of diagnostics include patients, clinic staff, laboratory managers, ministries of health, NGOs, regulators and suppliers. If we do not take the perspective of all users into consideration, we risk that these technologies do not fit their intended use setting, cannot be made to work and scaled up, are not utilized or not accessible for those in need. User perspectives on new diagnostics, their preferences and values, as well as their experiences with existing diagnostic systems, are important to take into account during WHO decision-making on new diagnostics, including guideline development and policymaking. Feedback from representatives of key stakeholders groups (including patients, health professionals and programme managers) is important.

Studies generating this kind of data are often qualitative in nature (i.e. they focus on meanings that people bring to a phenomena and how they act upon it). Qualitative studies use targeted sampling methods to capture diagnostic experiences across a range of users, diseases, tests and diagnostic settings (Engel et al., 2017; Engel et al., 2015; Engel et al., 2018; McDowell & Pai, 2016; McDowell et al., 2018; Miller, Parkhurst, Peckham, & Singh, 2012; Squire et al., 2005; Yellappa et al., 2017). They are an ideal method for making sense of user experiences with and perspectives on diagnostic tools within "real-world" situations because they avoid placing assumptions about what these tools are expected to accomplish at the outset (e.g., that a test is easy to use). By involving users (e.g., through interviews, usability tests, ethnographies and user feedback), qualitative studies can support decision-making on diagnostics and offer concrete insights into users' values and preferences, as well as acceptability and feasibility of new diagnostics in intended use settings. Such data will also point out important considerations for scale-up.

In December 2019, the World Health Organization will be updating the policy around molecular based tests for diagnosing TB and resistance to rifampicin, particularly looking at Xpert MTB/RIF, Xpert Ultra (both Cepheid) and preliminary data on TrueNat MTB/Rif (Molbio). To inform those discussions, the WHO has commissioned a study into the perspectives, preferences, and experiences of users of Xpert (including TB survivors, health professionals, and programme managers). To this end, we conducted a qualitative study with participants in Ukraine, Uganda, Pakistan and South Africa. We interviewed clinicians, laboratory staff, programme officers, TB survivors, and patient advocates with the aim to understand their experiences of using Xpert and diagnosing TB using molecular diagnostics more generally and to contextualize users' preferences.

This study is exploratory in nature and part of an ongoing inquiry into user perspectives of new TB diagnostics. More, in-depth ethnographic research on the ground is warranted to better understand perspectives and practices of different users including TB survivors, patients and their caregivers.

2. Methodology

In October and November 2019, NE and MW conducted semi-structured interviews with 23 Xpert users (including clinicians, programme officers, laboratory manager and technicians, TB survivors and patient advocates) in Ukraine, Uganda, Pakistan and South Africa. These countries were selected based on the fact that they have ordered large quantities of Xpert Ultra cartridges and are located in different geographical regions. Due to the short timeframe, participants were purposively sampled and approached based on convenience through personal contacts and colleagues. All interviews were conducted via the phone and in English. Four of the six interviews with participants in Ukraine were conducted with the help of a translator. We asked for the testing and treatment experiences as well as experiences on interaction between providers and patients to contextualize users' preferences about a new diagnostic. Topics discussed included: current approach to diagnosing TB using molecular assays including specific challenges; experiences with using Xpert MTB/RIF and Xpert Ultra specifically, including details on steps taken in the diagnostic process, determining eligibility, interpreting results and treatment initiation as well as challenges and benefits; ways of interacting with patients about Xpert; overall usefulness; the impact of Xpert including on access, equity and feasibility; and the current policy context. Xpert MTB/RIF is most widely in use and challenges to implementation have also been partially published (Albert et al., 2016; Clouse et al., 2012; Engel, et al., 2015; Hanrahan et al., 2015). The conversations therefore mostly focused on the experiences of using Xpert MTB/RIF and where in use also zoomed in on specific differences experienced with Xpert Ultra.
Interviews were audio-recorded, transcribed by MW, and coded by NE in NVivo qualitative data analysis software. We each wrote memos on different topics, discussed these and collated them into themes which we present below. Professional roles are used to mask study participants' identity. Because the interviews were conducted by the phone, it was not possible to triangulate interview data with other evidence commonly collected through ethnographic approaches (such as multiple interviews and informal conversations at the same facility, observations or site visits). This warrants more in-depth and on the ground research with face to face interviews to understand all user perspectives and practices of diagnosing TB in PLHIV.

Ethics
This study was approved by UMREC, the ethical review board of Maastricht University. Study participants were emailed an information sheet explaining the objectives of the study and an informed consent form which they signed prior to participation.

Table 1 Participants overview per country

	Ukraine	Uganda	Pakistan	South Africa
Clinician	3	-	1	1
TB survivor/ advocate	1	1	-	1
Laboratory manager/technician	1	2	2	2
Programme officer	1	1	5	1
Total	**6**	**4**	**8**	**5**

3. Results:

Below we discuss the results for current use of Xpert separately for the four countries and then discuss overarching themes that emerged from the interviews across the different countries.

Current use of Xpert MTB/RIF and Xpert Ultra in Ukraine, Uganda, South Africa and Pakistan
Ukraine started using Xpert MTB/RIF in 2012. Presently, the machines are placed in the labs of some primary and tertiary hospitals, as well as AIDS centers and penitentiaries. For facilities with access to

the machine, it is the first line diagnostic test for people with presumptive TB, otherwise sputum microscopy is still used. Sputum is the main specimen tested on Xpert MTB/RIF. While a clinician noted that it is very difficult to convince laboratory staff to run extrapulmonary samples (ID14), it seems that they are indeed being tested in higher levels of the system such as tertiary facilities and BSL 3 labs (ID16 programme officer & ID19 laboratory manager). It was however specified that in these laboratories extrapulmonary samples – namely urine, cerebrospinal fluid (CSF), pleural fluid, lymph node aspirate, and feces -- are tested on Xpert Ultra exclusively (ID19 laboratory manager). From our interviews, the introduction and scale-up of Xpert Ultra in Ukraine is otherwise unclear.
Like Ukraine, Uganda introduced Xpert MTB/RIF into the TB program in 2012. Where available, it is currently being used as the first line test for people with presumptive TB, especially in facilities located in districts with a high TB burden. Smear microscopy is used as a pre-screening tool for Xpert MTB/RIF, or as the main diagnostic test if Xpert MTB/RIF is unavailable. The main specimen run on the platform is sputum, but other extrapulmonary samples have been used occasionally and in research settings. Xpert Ultra cartridges have been rolled out with plans to completely phase out Xpert MTB/RIF (ID1 laboratory manager).
South Africa was an early adopter of Xpert MTB/RIF as a first line test for all people with presumptive TB (WHO, 2010),. and by 2018, it was the only country using Xpert Ultra as an initial diagnostic test. Presently, Xpert MTB/RIF cartridges have been completely replaced by Xpert Ultra cartridges. Sputum for adults and gastric specimens from children are the main pulmonary specimens used, while CSF and lymph node aspirate are the main extrapulmonary specimens. The machines are located in labs operated by the National Health Laboratory Service across various levels of the health system.
Of the countries represented in this study, Pakistan is the only one that does not currently offer Xpert MTB/RIF as a first line test for all people with presumptive TB owing to resource limitations. Instead, it is the initial test for presumed TB in children, drug resistant (DR)-TB, extrapulmonary TB (EPTB), and immunocompromised individuals. For everyone else, it is used as a follow on to abnormal chest x-rays and/or positive smear microscopy. According to programme officers, the National TB Program (NTP) is set to change the guidelines to expand the eligibility criteria for Xpert MTB/RIF in the near future (ID3, ID5, ID10). The machines are typically housed by district hospital labs, with some in lower level health facilities in higher burdened districts. Like the countries above, sputum is the main specimen used on Xpert MTB/RIF, with the occasional use of EPTB specimen such as CSF, stool, and urine. According to an NTB laboratory advisor, there are no plans to introduce Ultra in the near future due the cost and the shorter shelf life of the cartridges (ID3).
Where available, most of the providers among our study participants from these four countries used LPA or culture-based drug susceptibility testing (DST) as a follow-on test when rifampicin resistance was detected by Xpert. For those testing EPTB samples with Xpert, many understood the sensitivity of the test to be suboptimal, but nonetheless appreciated the ability to identify a few more patients than with conventional methods. In fact, if an EPTB sample tested positive on Xpert MTB/RIF or Xpert Ultra, it was often perceived as a true positive and followed by treatment initiation. Negative results however were not deemed as an indication of the absence of TB and were therefore accompanied by further investigations.

Xpert has helped to improve the diagnosis of drug-resistant TB
The participants we spoke to assign the greatest value to the ability of diagnosing drug-resistant TB with Xpert.

> *"Without Xpert it would not have been possible to put so many patients on second line treatment. That is for sure."* (ID10 program officer, Pakistan)

According to a clinician in South Africa, molecular testing has revolutionized the TB program in South Africa:

"I can tell you this, today you can be a TB patient suspect, if your sputum is positive for TB, in five days we can tell you with a 90% probability of accuracy, you have drug susceptible TB or drug resistant TB. That never happened in the past."(ID11).

According to clinicians and a laboratory manager in Ukraine, Xpert has increased case identification of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB in Ukraine and allowed faster treatment initiation. Clinicians do not need to wait for culture results to initiate treatment, because most of their RIF+ patients do have either MDR or XDR since there are only very few patients with mono- or polyresistance (ID12 clinician, ID19 laboratory manager, ID20 clinician). Also, clinicians are able to separate the patients physically into RIF+ and RIF- to prevent transmission (ID16 programme officer). For Ukraine a programme officer at the Ministry of Health mentions that the number of TB diagnoses has gone up too thanks to Xpert (ID16). A Ukrainian TB survivor highlights how the increasing use of Xpert means patients are diagnosed before they are hospitalized (in case of MDR-TB) instead of the other way around. This is crucial as hospitalization can cause much suffering especially if hospitalization turns out not be justified (ID21). Even though the TB survivor was diagnosed in a hospital with Xpert on site, the doctor at that time did not use the Xpert for her diagnosis and therefore could not determine whether she had susceptible or drug-resistant TB. It is unclear why the test was not used, one possible reason being stock outs. The hospitalization meant the TB survivor was separated from her child who was hospitalized in another hospital for four months from which both of them suffered psychologically (both needed long-term psychological counseling afterwards). What is more, the survivor lost her job due to hospitalization. If she had been diagnosed as susceptible at that time she would not have been hospitalized (ID21). This example shows the powerful impact on patients if they cannot access testing for drug resistance through Xpert.

In Uganda, according to a laboratory manager, Xpert increased the number of patients with DR TB diagnosed, so now the problem has moved from being one of diagnosis to one of the logistics and time of treatment initiation (ID1).

A clinician in South Africa reflects on the difference Xpert made for his ability to not only diagnose MDR-TB but also to treat MDR-TB which was not possible without a test:

"..[many years ago], When I have to wait sometimes six months to have someone diagnosed with drug resistant TB. When actually almost there were no policies about it; and when you sometimes suspect it and you say but this patient may have this, and then you go to pharmaceutical and say this patient have drug resistant TB, I need this and this, and then they say 'ok show me the test'. And you don't have a test, so you cannot treat the patient." (ID11)

Mixed experiences with establishing the impact on diagnosing EPTB, case notification, and loss to follow up

The participants we talked to mentioned the ability of Xpert to diagnose different specimen types positively. However, it is less clear from the data how often these options can be used and what their impact is. A Ugandan activist mentions, in particular, the positive impact of Xpert on diagnosing patients co-infected with HIV (ID15). The ability to use different sample types allows clinicians to try out different samples and once in a while in Ugandan routine care this is how a patient with EPTB is diagnosed (ID1 & 2 laboratory technicians). Specifically for Xpert Ultra, laboratory technicians in Uganda mention that with Xpert Ultra, they get more positive results (ID1,2). A programme officer in Pakistan observed an increased yield with CSF due to Xpert Ultra and more trace in processed, concentrated specimens (ID10).

In Pakistan, programme officers involved in an active case finding programme where Xpert is combined with upfront chest X-ray, mention that besides more and quicker detection of RIF resistance, Xpert is able to detect more patients with low bacillary load (ID 5 and 6).

Programme officers of the NTP, however, point out that the notification rate for susceptible TB is high and static and does not seem to increase with Xpert (ID3, 10), since so many patients are

clinically diagnosed even though the rate of bacteriological confirmed TB increased a bit (ID10). The notification rate for pediatric TB and EPTB are not affected by Xpert either, because the specimens are difficult to get with current facilities and capacities (ID10). A pediatrician confirms that the bacteriological confirmed percentage of patients has gone up with Xpert, but the majority of children are still diagnosed clinically because they cannot produce sputum (ID18). For EPTB, very few public sector facilities in Pakistan have laboratory services offering histopathology. These services are common in the private sector while bacteriology is available in the public sector. This means that clinicians in Pakistan need to make a choice for each specimen between histopathology and bacteriology. If sent to histopathology in the private sector, the sample will not make it back onto an Xpert machine in the public sector (ID10). The diagnostic test available in the public sector is unable to bridge the disconnect in laboratory services.

Consequently, for a programme officer of the NTP, the value of rolling out Xpert lies mainly in diagnosing RIF-resistant TB and MDR, but not necessarily in diagnosing susceptible TB. Given the limited impact on case notification, the challenges with maintaining the machines and the high cost of cartridges that prohibit simplified algorithms (for instance Xpert for all presumptive TB patients), the programme officer is cautious to advocate for more widespread use of Xpert (ID10).

According to our participants, and somewhat unexpectedly, in Pakistan, South Africa and Uganda, the introduction of Xpert also increased the turn-around time from the 24hrs for smear microscopy mainly due to the challenges of utilization discussed below (ID1, ID9 laboratory managers, ID10 programme officer, ID13 clinician). This could potentially affect case notification as well, because an increase in TAT of Xpert comes with higher risk of loss to follow-up. However, in Pakistan a percentage of those lost to follow-ups might be intentional, because patients from the private sector access free microscopy and Xpert testing in the public sector but are not necessarily notified to the government (ID10).

Xpert's convenience and automation has eased laboratory work

Compared to smear microscopy, users generally value the automation, convenience, higher biosafety levels and lesser human involvement that Xpert offers (ID1, ID2, ID19, ID22, ID23 laboratory managers and technicians, ID16, ID17 programme officers, ID18, ID20 clinicians). The fact that it is a closed system with walk away time during the incubation (15') and machine run time (90') where laboratory technicians can do other testing in between was mentioned as well (ID1&2). Specifically for Xpert Ultra, the fact that Xpert Ultra takes less time can be helpful in some situations (for instance an active case finding setting with high throughput) (ID6 programme officer). As such, Xpert eased the work for laboratory technicians, adding a level of relief from reading hundreds of slides (ID9) as well as reducing the room for errors (ID22). This convenience can have undesired consequences for sites where Xpert is not available: In Uganda, even though Xpert is not available at every site, health workers are reluctant to use microscopy and so monitoring has gone down. A laboratory manager speculates that health workers would prefer transferring responsibility of testing to an Xpert site or might be too lazy to do the more cumbersome sputum microscopy testing (ID1) or might struggle with stockout of supplies.

Clinicians' confidence in Xpert results

Laboratory technicians in Uganda emphasize how Xpert impacted the clinicians' confidence in its results due to its increased sensitivity and

> "*since this is a fully automated system, the clinicians have that confidence that this is actually,.. could be the true status of this patient in terms of TB*" (ID1 & 2).

A programme officer in Pakistan mentions how a negative test is not a barrier to diagnose a patient with TB (ID10). In other settings, physicians have become more hesitant to diagnose TB clinically. Another programme officer from Pakistan observed that if it's a negative result on Xpert the level of suspicion among the clinicians for TB has gone down drastically; at times its good but they might

miss patients then clinically (ID5). That is also the case when sputum microscopy is used concurrently.

> "*Especially for follow-up of course, but even at baseline they [clinicians) continue to use it [sputum microscopy] as an adjunct. But their decisions are I think primarily that wherever there is Xpert, the Xpert results pretty much are their diagnoses.*" (ID6)

This is also true for a pediatrician practicing in a hospital in Pakistan who immediately places children and adults with positive Xpert results on treatment. She is confident in the true positive nature of the test also because children who turn positive on Xpert are mainly those already admitted and who are very sick and not those being sent from the outpatient department (ID18).

Trace: "it has not helped in decision making" – ID10 program laboratory advisor

Xpert MTB/RIF Ultra's additional category of trace was often said to be accompanied by uncertainty. However, a clinician noted that the trace category is a welcomed indication of a test that can detect very low levels of TB (ID11). Although most professionals that were interviewed said the policy is to follow the WHO guidelines of re-testing positive trace samples on fresh specimen of those who are not PLHIV, children or suspected EPTB, actual laboratory and clinical management of trace results was rarely as straightforward. For example, a program officer from Uganda noted the difficulty in obtaining a fresh specimen to run a second test:

> *"[trace results] are causing quite a commotion for health workers. They are not following the algorithm… [because] by the time most of these patients get back these results, they have already left the health facility*" (ID4).

Others reported similar challenges with getting a second result, due to loss-to-follow up (ID16, laboratory technician; ID1, laboratory manager; ID6 program manager; ID8, clinician) or the patient's inability to produce sputum again if they have since been initiated on treatment (ID13, clinician).

If a second confirmatory test is actually done, professionals reported being confused by the results, specifically if the second result is also a trace or negative (ID6, program manager; ID18, clinician). For example, one clinician noted that if trace positives are deemed as true positives, a second negative result may cause confusion on whether to start treatment (ID18, clinician). Furthermore, an additional test raised questions for some laboratory professionals on which of the two results to report to the clinician. For example, a program officer in Pakistan noted

> "*our experiences with the Xpert Ultra, sometimes because the guideline is that if we have a trace result you repeat the result, and sometimes when we repeat the results we get MTB not detected….what should we report because the repeat testing has actually not helped us*" (ID10).

This uncertainty may be why the number of reports a clinician receives varies. A clinician in Pakistan reported only receiving the second report (ID18 clinician) whiles laboratory staff in Pakistan stated that they provide both reports to clinicians (ID22 laboratory supervisor). And if the policy is to perform culture as a confirmatory test for trace results, the long turnaround time may present challenges for clinical decision making (ID2 laboratory manager, ID8 laboratory manager).

Additionally, discordant results between Ultra and phenotypic tests like culture were also reported to raise questions on the validity of the trace result (ID2 laboratory manager; ID10 programme officer; ID8 laboratory manager).

If at all the clinician receives one or both trace results, it was often reported that the result then becomes a small piece of a big decision-making puzzle (ID1, ID16, ID8, ID23, ID19 laboratory managers). As a clinician in South Africa notes,

> *"..actually when you have one of these results, that actually is not telling you that much. What you need to do is go back to the drawing board and say ok, what is the probably that this is really TB….what is the clinical condition of the patient…the TB history of the patient….you need to put all this together and then make a decision…you have in front of you a patient, not a laboratory result"* (ID11).

While some clinicians didn't seem to have a problem with having to use a more intensive evaluation process for trace patients (ID11; ID20), others cautioned that not everyone would have the expertise to do such, especially frontline health workers working in peripheral facilities (ID6 program manager; ID11 clinician; ID18 clinician). As a program manager notes:

> *"It's the larger sort of doctors and things where I think for them interpreting these results, and understanding what needs to be done, certainly requires a lot of capacity building and training because to them this is still not very clear, on how these different results and the two results need to be sort of tallied and what it entails and what protocol or process they need to follow depending on the results. So, if the first one is trace and the second is trace, what are the next steps, what are they looking at, what do they need to do if the second one is negative then what do they need to do..." (ID6)*

A pediatrician also cautioned about the specific challenges trace results might present for those diagnosing TB in children. She notes that diagnosing TB in children is already difficult, and apprehension of the validity of a trace result may only add to the confusion (ID18).

Lastly, it often came up that the absence of information on rifampicin resistance slowed down clinical decision-making as confirmatory DST and LPAs needed to be done (ID 2 laboratory manager; ID10 programme officer; ID11, ID13, ID18 clinicians; ID22 laboratory manager), presenting unique challenges for high-burden DR, MDR and XDR TB settings.

Weighing the trade-off overtreatment vs missed diagnosis

When asked whether it is better to detect more TB, possibly at the risk of falsely doing so, than to miss true cases, some respondents were in favor of the former (ID11 clinician, ID18 clinician). As noted by a clinician from South Africa:

> "*A balance between potential harm and overtreating TB patients must be given… the balance is more to the risk of trying to use molecular testing in order to diagnose more TB, even when we know we may over diagnose too much, but the benefit from the individual [patient] point of view of detecting tuberculosis early, and the benefit to cut off the transmission early, probably outweigh the risk of this."* (ID11).

To some, managing the side-effects of anti-TB treatment is a better outcome than missing a true TB case (ID11, ID18 clinicians). This was echoed in a study on user perceptions of Xpert Ultra during decision-making workshops on the transition from use of Xpert MTB/RIF to diagnose TB to Xpert Ultra in Swaziland and Kenya by the authors of this report (Mwaura et al., under review). The participants of this particular study attending these stakeholder meetings, including clinicians, laboratory technicians, TB program coordinators, patient advocates and NGO representatives felt the harm to the individual and community of a missed diagnosis outweighed that caused by a false TB positive.

Conversely, overtreatment was not perceived to be without faults. One clinician noted the damage false positives would do to community confidence in the healthcare system (ID18), a sentiment also brought up by health workers in Swaziland (Mwaura, et al., under review).

But ultimately, for most of the respondents, managing this tradeoff between overtreatment versus missed diagnosis is really on a case-by-case basis that is up to the discretion and expertise of clinicians (ID1 laboratory manager, ID10, ID6 program managers, ID11, ID12, ID20 clinicians).

Discordant results and tie breaker

When discussing challenges with using Xpert, participants of the study often noted the difficulties of interpreting discordant results. Specifically, the RIF-resistant information from Xpert was observed to be unreliable (when compared to LPA and culture) when MTB was detected at very low levels (ID2, ID22 laboratory managers, ID10 programme officer). While clinicians acknowledged that Xpert is only one piece in the diagnostic portfolio that informs clinical decisions (ID11, ID20), the caveat to

this was that it takes training, experience and expertise to understand and contextualize conflicting information regarding your patient's status (ID6 program officer, ID11 clinician). This may be further exacerbated by questions around which test is truly the gold standard, as culture may not always be accurate due to usage of poor-quality specimen (ID1, ID8 laboratory managers). As such, a laboratory manager hoped for the development of a reliable and valid rapid diagnostic test that could serve as a tie breaker when the results of different tests or different results of the same test were not consistent (ID2).

Xpert and previously-treated patients

Since Xpert relies on molecular detection of genetic information (MTB-complex DNA), there is no clear indication of whether detected mycobacterium is alive or dead. This may present unique challenges when testing individuals who have been previously treated for TB, especially when the parameters for 'previously treated' are not clearly defined. For some, previously treated are those who completed their treatment six months ago (ID11 clinician), while for others it was one year (ID2 laboratory technician) and two years (ID20 clinician). And if the parameters are defined, establishing a thorough TB history was said to be uncommon, thus increasing the risk of wrongly initiating treatment in an Xpert positive case (ID2 laboratory technician). But if a thorough TB history is established and a clinician opts to not initiate treatment following a positive Xpert result, it was noted that this may inaccurately lower national statistics of the percentage of bacteriologically confirmed cases that are initiated on treatment (ID11 clinician). As such, there need to be clear parameters of how to define this patient category, handle their Xpert results, and accurately capture outcomes in national databases.

Counseling and interaction with patients needs more focus and care

Patients are usually not informed about the type of diagnostic test that will be run, and results are usually communicated through nurses or clinicians. Counseling is not consistently available according to our participants. A program officer in Uganda laments this lack of skilled counselors: "*We have nurses who worked as counsellors but they are not good at counselling. And even worse most of the sites do not have any counsellors specified for TB.*" (ID4) A TB survivor in Ukraine highlights that patients are only informed about the waiting time for results with regard to diagnostics:

> *"So when they provide counseling to the patients they usually say how long the patients have to wait for the test results and in principle that's all. (…) But nobody explains to the patients anything about the form of tests or molecular tests" (ID21).*

A TB survivor in South Africa received no counseling alongside his first diagnosis of MDR-TB and doubted the initial diagnosis (ID7). Considerable cost and time is needed for patients to come to the clinic for medication, follow-up testing and check-up appointments with the doctors all at different moments in time (ID7). The TB survivor explains how simply explaining to patients the current status of diagnostic technology and that the providers would continue testing the sputum for resistance since the current test is not conclusive would aid in justifying even more waiting time and investment of cost and time and allowing the patient to understand what is happening.

> *..you wait for the certain period, the results come back as saying you have a TB. They not sure which TB it is, and then for so long they will give you a normal TB treatment. After a month, or weeks, or three weeks, they find out no, you are not normal TB, you are MDR TB, you must change the treatment. So, we don't know exactly where this come from. So, as a patient you won't understand. If at least in the first place they told you, we have a problem with the diagnostic tool. So we are gonna test you as long as the results come back you are having the TB but we are still going to continue with the test at the laboratory to check what kind of TB do you have." (ID7)*

The relationship with a nurse or DOT provider is focused around treatment and while South Africa has introduced counselors for MDR and XDR-TB patients, the TB survivor laments that these counselors have been trained on HIV and are not sufficiently specific about DR-TB:

> *" those counsellors they are the people been trained to counsel people with HIV, they are not specific about DR TB. They are just touching the baseline of XDR or MDR, they don't have the full information about it."(ID7)*

A laboratory manager and technician in Uganda found in their research studies that involving a social worker at the clinic who does patient education, counseling, symptom screen, supports patients with sample provision, bringing samples to the lab, knows when test results are available and has repeated contact with patients in their own language throughout their diagnostic journeys has a big impact on patient satisfaction. It has improved patient retention and provision of additional samples if needed (ID1.ID2).

Feasibility and utilization

According to our participants, underutilization of Xpert machines still poses a problem for many sites in Pakistan, Uganda and also Ukraine compounded by the challenges of delays due to sample transport, module break down, stock out of cartridges or complicated diagnostic algorithms (ID1 laboratory manager, ID2 laboratory technician, ID4, ID10 programme officers, ID12 clinician, ID15 activist). Programme officers, laboratory technicians and activists in Uganda and Pakistan mentioned how clinicians are reluctant to use Xpert as baseline, as a result of these challenges (ID1, 4, 22, 23, 15). Instead sticking with the, in their eyes, more cost-effective and easily available smear microscopy (ID22, 23, 15).

What is more, programme officers in Pakistan and Ukraine voiced concerns about the cost and the sustainability of donor dependency for Xpert testing in the longterm (ID3, ID5, ID6, ID16). Below, we discuss the impact of sample integrity and transport, stock out, maintenance and repair and simplicity of diagnostic algorithms on feasibility and utilization in more detail.

<u>Sample integrity and transport</u>

As the cost implications of having an Xpert machine in every TB testing site would be enormous, the four countries each rely on inter-facility sample transportation. Participants from Ukraine, Pakistan, and Uganda noted various challenges with their current transportation system. A laboratory manager from Uganda was concerned that while theirs has improved access of patients to Xpert testing, it is still not as efficient and can cause TATs of two days to two weeks (ID1). This may result in health workers opting to use smear microscopy in their sites instead of risking delayed treatment initiation (ID1 laboratory manager, ID4 programme officer). Similarly in Ukraine, a programme manager highlighted specimen transportation as the major challenge (ID16). A clinician noted that even though TAT is usually one day, their once-a-week sample transportation system still delayed initiation of treatment (ID12). In Pakistan, a program officer observed that there is currently no established system for the transportation of any clinical samples, and if the TB program managed to develop one, it would actually be a pathfinder for the country (ID10). This results in limited access to Xpert for patients visiting facilities without the machine.

If samples are to be transported, upholding their integrity before or during transportation was a cause for concern for some (ID1 laboratory manager, ID4 program officer and ID18 clinician). For example a program officer noted that despite triple packing the samples during transportation, "*its an issue…especially at the storage before they take that sample…because there are no good storage facilities and most of these health centers where they don't have a gene Xpert are very low health centers, not like high volume health centers*" (ID4). On the flip side, a South African clinician was of the position that the quality of the sample is more important than its conservation, since Xpert relies of the presence of mycobacteria DNA, whether alive or dead (ID11).

Stock out of cartridges has enormous impact on utilization

The users in Ukraine and Uganda we spoke to mentioned supply and specifically stock-outs of cartridges as a major challenge to utilizing Xpert (ID1 laboratory manager, ID2 laboratory technician, ID4 program officer, ID12 clinician, ID15 activist, ID21 TB survivor). A clinician in Ukraine is asked send more or less patients for testing to the laboratory depending on the supply (ID12). In Uganda, according to a laboratory manager stock-out of cartridges happen twice a year (ID1). A program officer confirms that because cartridges are very expensive the country has not been able to buy them at expected levels and therefore supply can be less than what health facilities order (ID4). The impact of stock-out and disruption of testing on laboratory work and clinical practice can be quite large. If Xpert has replaced microscopy as the first-line test, the equipment necessary to run smear microscopy might not be available anymore, as a laboratory manager from Uganda recounts, leading to chaos, delays and unsatisfied patients, during a shortage of Xpert cartridges in Kampala:

> *", .. probably even the slides may not be there, so in your order menu [for your laboratory] you may not now include things for staining, I mean like you need so many slides, you need…so when it [Xpert] breaks down and the process of acquiring all this [supply for smears] takes a little bit of time, so you've got a gap, a huge gap. And patients keep pouring more samples. Recently we had like a cartridge shortage around Kampala, all the fridges were full with samples, and patients were not happy. You know, they were like 'these guys are not working enough!' And it was a total kind of chaos. (…) there is actually a big impact because Xpert has actually become like a point of care test. Microscopy is mostly done on follow-ups in most health centers so there is a big impact when there is that shortage."* (ID1)

Stock-outs of cartridges in combination with power shortages deter clinicians from ordering Xpert and instead relying on smear microscopy according to an advocate (ID15). A program officer explains that due to the cartridge shortage, Uganda had to change their algorithm: instead of testing all presumptive TB patients on Xpert (their current policy), the NTP had to limit direct Xpert testing to highly suspected patients; for example contacts of MDR-TB patients, HIV positive patients, diabetic patients, minors or those with risky occupation (ID4). This complication of the algorithm might negatively affect utilization further.

Participants in both Uganda and Ukraine locate the issue with stock outs to government funding. A TB survivor in Ukraine explains that stock outs happen due to insufficient allocation of funds by the government:

> *"..even the largest hospitals and clinics that have GeneXpert which was bought for the donor money… unfortunately the State Budget did not allocate funds to purchase these cartridges and there is often a situation when they have the equipment but do not have cartridges.*" (ID21)

According to an activist, laboratory services in Uganda are up to 99% donor funded. When donors procured Xpert with the understanding that the government would purchase cartridges and service these machines, it did not happen because there is no specified budget line for laboratories. According to this participant, the laboratory has been an afterthought, with insufficient funding and attention to policy implementation or training, and therefore poor adherence to diagnostic algorithms in general (ID15).

Bottleneck maintenance and repair: High workload and poor supervision in combination with infrastructure and environmental conditions cause frequent module breakdown

The respondents we spoke to reported frequent module failures (ID1 laboratory manager, ID2 laboratory technician, ID5, ID6 program managers, ID8, ID9 laboratory managers, ID10 program officer)). A laboratory manager of a sub-district level clinic in Uganda illustrates how module break down prolongs turn-around time, causing backlogs, higher workload and contributes to underutilization:

> *"16 [samples per day] is usually on the ideal end, because there are cases whereby maybe (..) run 4 samples and one [module] consistently, because there is an error, like you do not have a valid result at the end of the day. So you have to redo it and the other ones have to keep waiting until maybe you get a valid result." (ID2)*

A laboratory manager in Uganda explains that high workload in combination with infrastructure and environmental conditions cause frequent module breakdown; if sites are busy and infrastructure is poor depending on where exactly the platforms are placed, machines break down:

> *"we realized that if the laboratory techs consistently maintained the Xpert machines according to the manufacturer recommendations, we would actually substantially reduce the error rate. And so, yeah, those errors occur in places where workloads are high, people do not have adequate time to do daily maintenance or regular maintenance. We have also realized that the recommendations probably are not environment based. In areas where you find there's a lot of dust, which is of course common in our settings here, instead of doing probably a weekly maintenance schedule, you might need to do it more frequently" (ID1)*

The presence of local CEPHEID agents who do annual calibration and trouble shooting seems key to maintenance In Ukraine and Pakistan (ID19 laboratory manager and ID20 clinician), but even then if staff fails to do daily/weekly maintenance, modules still fail frequently. Laboratory managers in South Africa emphasized that continuous training to do proper maintenance, especially in a context of high staff turnover, and dusting of modules with a specific brush is required to keep error rates low (ID8 and9). In Pakistan, programme officers found that maintenance by technical staff is often poor despite training, due to insufficient supervision and follow-up (ID10, 6). According to one of the programme officers, especially the monthly maintenance, which involves opening up the machines and cleaning the filters, should be done by specific staff responsible for just maintenance tasks across several sites. They experienced many errors when the local laboratory technicians did the maintenance work, not putting things back into the right place or not working carefully enough (ID 6).

What is more, repair and replacement of modules takes too long according to programme officers (ID3, 10, 5, 6). A programme officer in Uganda highlights the delay caused by the reliance on one specific supplier and repair person:

> *"we have only one person who repairs those machines, and sometimes Cepheid delays to come and repair the machines. So, the health workers sometimes they take some time as they are waiting for Cepheid to come and repair the machines, and we are missing our patients"(ID4)*

A programme officer in Pakistan recounts how replacement of a module before they had a local Cepheid representative would take 4-6 months. Trying to avoid these delays, the programme had bought 100 modules extra as stock to fall back to. But in order to replace faulty modules in the machines with their own stock they needed the accordance of Cepheid which generally took 1-2 weeks (ID10). As with cartridge stock-outs, the impact of module failure and slow repair on laboratory and clinical work and ultimately utilization and access is enormous and long felt. A programme officer in Pakistan illustrates this impact on work- and patient flow:

> *"It is difficult to maintain a workflow or a practice (..) if the services go out of order. (…) because the physicians start referring patients for testing and then one odd day they would learn that machine, all four modules, is out of order, and there is no services available, there is disruption of services. And so the next time when you have your modules functioning, you have to restart your training, and practices, telling people that ok now the machine is ok you can refer patients back. So your flow of the patient is broken."(ID10)*

<u>Simplicity of diagnostic algorithm is essential for feasibility and utilization, but crucially dependent on cost and supplies</u>

The challenges of cost and supplies related to Xpert can complicate diagnostic algorithms with undesired consequences for utilization of Xpert: health workers then stick to a simpler algorithm

involving the old diagnostic and avoid ordering Xpert. According to a program officer in Uganda, Uganda's simple algorithm enhances feasibility of Xpert:

> *"Feasibility is not a problem, as I am saying, our Xpert algorithm is quite simple, it says all presumptive patients send them to Gene Xpert, (…) So, in terms of feasibility I think its ok, it can be done very fast….it doesn't have a lot of do this, do this, do this,.."*(ID4)

The program officer stresses how this simplicity is complicated during cartridge stock-out, when health workers need to decide which samples should make it onto an Xpert machine and which not (ID4).

A program officer in Pakistan explains how Xpert in the government clinics is currently underutilized partly because of the problem of module break-down but also because of the rather complicated diagnostic algorithm (ID10). The original diagnostic algorithm in Pakistan suggested to use Xpert as an upfront test for all patients with a history of previous treatment, context of MDR, pediatric patients, extrapulmonary patients, and HIV positive patients and as a follow-on to microscopy for the smear-positive patients. This changed when the country increased the number of Xpert machines. Now all patients who already had an X-ray done and have abnormal X-ray can do upfront Xpert testing. But the programme officer is unsure if these algorithms are followed. While the testing for rifampicin has increased, it is hard to determine from the dashboard connected to Xpert whether patients are directly tested with Xpert or first given a sputum microscopy. She suspects that clinicians keep sending for smear microscopy first, knowing that if the smear is positive the patient will anyway be sent for Xpert testing; also because Xrays need to be paid out of pocket:

> "*Because for some of the places it's still MTB positive is very high, and it looks as if they are still following the DR testing as follow on to microscopy*" (ID10).

The program officer provides an example of a chest hospital that conducts chest X-rays routinely on their patients but had not been utilizing their Xpert machine much. When that hospital switched to a (for their particular situation) simpler algorithm (i.e. everybody with abnormal chest X-ray gets Xpert) their use of the installed Xpert and their numbers of bacteriologically confirmed pulmonary TB went up dramatically. But according to the program officer, the NTP in Pakistan is hesitant to support utilization by promoting simpler algorithms for facilities that have access to sufficient cartridges (test all presumptive TB patients for instance), because of strong advocacy voices within the country that demand equal access to Xpert testing for everybody across health facilities. However, testing all presumptive TB patients across the country is not possible because of lack of resources and poor/lack of specimen transport across the country (ID10).

A diagnostic algorithm involving Xpert that is simple to follow for the particular facility with its own resource situation seems essential for feasibility and access, but is crucially linked to cost and supplies (f.i. chest X-ray and cartridges). It also shows how in many places, cost and supplies mean that screening with Xpert is currently not a feasible option.

Transparency and accountability of the implementation process increase utilization and ultimately access to Xpert

In Uganda, donor supported high burden sites have intensified TB efforts. In contrast to government facilities, where according to an activist there is a lack of government employed and sufficiently trained staff to operate Xpert machines, international donors are supporting laboratory technicians on a project basis (ID15). What is more, while donor supported sites have cartridges available according to their needs and performance, government facilities are rationed according to the facility budget. According to an activist, the implications of the challenges with capacity, stock-outs, maintenance and power shortages for utilization are quite strong with the result that many clinicians are reluctant to use Xpert and most of the Xpert machines in government facilities are underutilized:

> *"In the government facilities where there are Gene Xpert machines, I think most of them are down actually, like there is no cartridge, they are not serviced and in most of those places where they are, like the regulation is not, the control is, like there is not so much control*

actually. To us, it's like they just procured and dumped equipment in some of these facilities without clear guidance and control over them." (ID15).

A solution could be to increase transparency and accountability of the agreements between donors and governments regarding introduction of new diagnostic technologies (ID15):

"I think it needs more civil society advocates involved in these processes, to ensure that we have social contracts with our government and we are able to follow them." (ID15).

The activist mentions an example of a clear and transparent memorandum of understanding on certain HIV diagnostics between PEPFAR and the government where civil society participated in the negotiations with the government. This enhances accountability and responsiveness from the government (ID15). Limited transparency makes it hard for civil society originations to push for more accountability. On a local facility level, each Ugandan health facilities in theory should have community boards with representatives from the communities. If these boards function and they understand the advantages these diagnostic platforms bring to communities, they can then demand the services, own machines and push those responsible for action (ID15). A TB survivor in Ukraine would similarly like to see improved social contracts between patients and TB programs to establish standards of counseling and to involve community representatives in decision-making around guidelines (ID21).

What users want for the future

Going forth, respondents are highly anticipating innovations that could improve the TB diagnostic landscape even further. For example, some hoped for cartridges that could provide more drug resistance information (ID11 clinician, ID12 clinician) or having LPA testing upfront available to determine resistance to isoniazid (ID11). Similarly, a point of care test that can rapidly detect TB in EPTB samples was suggested by a clinician:

" as a clinician I would like to have a test which is point of care test basically, which is able to detect TB from non-infectious samples like urine, like TB LAM or any other sample, irrespective of the site of tuberculosis in the body. So I think Gene Xpert the sensitivity is good for maybe sputum or maybe from the CSF but most extrapulmonary tuberculosis it is very difficult to rely on Gene Xpert to rule out tuberculosis" (ID13 South Africa).

Another clinician stated that instead of tailoring the test to available specimen, we need to find ways of obtaining better specimen on the different compartments in lung lesions (ID11).

Some participants imagined that a portable POC test would improve access to TB diagnostics by overcoming challenges around TAT, interfacility specimen transport and unstable power supply (ID1 laboratory manager and ID5 program manager), ultimately bringing the TB testing even closer to the community (ID15 activist). A TB survivor brings this to the point and compares it with shorter TATs for HIV testing and treatment: "*HIV you are tested today, you are diagnosed today, you get a treatment today. Why they can't do the same with the TB? Test me now, diagnose me now, I must get a treatment now*." (ID7),

Nonetheless, if interfacility transportation is still to be relied upon, lower level facilities need solutions to preserve specimen without cold chain as the infrastructural limitations that prevent these facilities from housing a Xpert machine may be the same limitations that prevent them from adequately preserving the specimen as it awaits transportation (ID1 laboratory manager).

4. Conclusion and recommendations user perspective Xpert testing

The results show that our participants assign great value to the ability of Xpert to improve the diagnosis of drug-resistant TB and conversely the impact on patients if they cannot access testing for drug resistance through Xpert. The impact on case notification and the value of Xpert for finding more TB was less clear owing to widespread clinical treatment, prolonged TATs and the challenges with feasibility and utilization of Xpert.

While access has improved, not everybody who needs it can access Xpert testing. Importantly, simple to use in the laboratory does not automatically translate into feasibility. Rather, feasibility of

Xpert testing depends on government commitment to ensure functioning infrastructure and power; supply of cartridges and functioning laboratory services; investment in expertise to handle (discordant) results; better repair services; staff with monitoring capacities; functioning sample transport; sustainable funding models and transparent donor agreements; and simple diagnostic algorithms. These aspects interact and reinforce each other determining utilization.
With regard to acceptability: while Xpert has eased laboratory work through convenience and automation, this preference for Xpert in the laboratory can have undesired consequences for monitoring through microscopy or for reverting back to microscopy when Xpert machines are down. While clinicians' confidence in Xpert results is rather high, the challenges with feasibility and utilization mean clinicians are at times deterred from ordering Xpert.
Below we discuss some of the results in more detail:

1. **Xpert is unable to bridge disconnects or lacking capacities in general laboratory services**. While participants valued the option to use other specimen than sputum, just having Xpert machines available in the public sector does not mean facilities and capacities exist to extract and make use of those specimen. For example, services for histopathology and bacteriology in Pakistan are disconnected and sending specimen to histopathology in the private sector, for instance, means the sample will not return to a public sector Xpert machine.
2. **Trace complicates decision-making**: laboratory and clinical management of trace results was rarely as straightforward. Study participants reported challenges with obtaining a second fresh sample when patients had left the facilities or had since been put on treatment and could not produce sputum as easily. If repeat tests are conducted after trace, they cause confusion when the second test is also trace or negative. Some laboratory managers are unsure which result to report and clinicians need expertise and experience to conduct more extensive evaluation for trace patients. This presents challenges for peripheral settings and where TATs of confirmatory tests (DST, LPA) slow down clinical decision-making.
3. **Discordant results of repeat tests and confirmatory tests can cause confusion around what should be considered gold standard**, particularly when specimen quality might be poor. Understanding and contextualizing discordant results require continuous training, experience and expertise.
4. **Establishing a thorough TB history of patients is uncommon and 'previously treated' defined differently** with implications for potential of false positives results through Xpert testing. Clear guidance is needed of how to define previously treated patients, how to handle their Xpert results, and accurately capture outcomes in national databases.
5. **The lack of trained counselors and of information provided to patients on diagnostics have negative implications** for their willingness to accept a diagnosis and invest time and money for clinic visits, follow up tests and treatment. Patients need better quality counseling by health workers to make it through diagnostic journeys and treatment, including information about diagnostic technology and considerations for follow-up testing.
6. **Persistent underutilization of Xpert machines is compounded by the challenges of delays due to sample transport, module break down, stock-out of cartridges or complicated diagnostic algorithms**. The presence of local CEPHEID agents is key for repair. But high workload and staff turnover, in combination with infrastructure and environmental conditions still cause frequent module breakdown and repair work can be slow or services deemed insufficient. The challenges of cartridge stock-out cause important delays and disruption of workflows leading to underutilization.
7. **Diagnostic algorithms that are simple to follow in a specific facility (f.i. test all those with presumptive TB) are more feasible and enhance utilization, but this simplicity is crucially dependent on cost and supplies**. Cartridge stock-outs or prohibitive costs can complicate diagnostic algorithms making them less feasible to follow further compounding underutilization. In Uganda Xpert testing eligibility criteria had to be temporarily restricted to certain patient groups due to cartridge shortages complicating the algorithm.

8. **Current donor agreements with governments** regarding introduction of new diagnostic technologies are not transparent enough **for civil society to be able to hold accountable and follow up**. Involving civil society in negotiating agreements and social contracts at national level and local facility levels can enhance accountability and responsiveness of governments leading to improved implementation processes and access to diagnostics.

Previous studies on Xpert have discussed the concerns around cost, slow policy uptake as well as underutilization of Xpert (Albert, et al., 2016; England, Masini, & Fajardo, 2019; Gidado et al., 2018). The current WHO recommendations, to run Xpert for all presumptive TB patients are not feasible due to the high cost and volumes of tests needed and explain why countries have a tendency to ration Xpert and limit access to high risk groups (England, et al., 2019; Pai & Furin, 2017). According to our study participants, this rationing complicates diagnostic algorithms, further decreasing feasibility and utilization of Xpert testing in already difficult circumstances.

Problems with maintenance, long repair and replacement times and insufficient service offered by CEPHEID were also found in the survey by England and colleagues and undermine the confidence programmes have in the affordability and sustainability of the technology (England, Masini, & Fajardo, 2019). Sustainable funding for networks of Xpert, sufficient cartridge supply and maintenance of machines is needed (England, et al., 2019), as well as investment in health system strengthening (Albert, et al., 2016), training and refresher training and policy sensitization with peripheral clinical providers regarding the rationale for prioritizing use of Xpert over smear microscopy (England, et al., 2019).
Our study also highlights the need..:

- for more investment in continuous training, experience and expertise to interpret discordant and trace results;
- to define previously treated patients, how to handle their Xpert results, and how to accurately capture outcomes in national databases;
- to simplify diagnostic algorithms adapted to the local situation while improving transportation systems to ensure access and utilization;
- to ensure trained counselors and comprehensive information on diagnostic technologies are available to patients;
- and finally, the urgent need for more transparency and involvement of civil society organizations and patient/community representatives to improve accountability mechanisms and implementation processes for new TB diagnostic technologies. This could help to monitor and address implementation challenges.

Acknowledgements
We would like to thank all participants for the time and insights they provided.

5. References

Albert, H., Nathavitharana, R. R., Isaacs, C., Pai, M., Denkinger, C. M., & Boehme, C. C. (2016). Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? [10.1183/13993003.00543-2016]. *European Respiratory Journal, 48*, 516-525.

Clouse, K., Page-Shipp, L., Dansey, H., Moatlhodi, B., Scott, L., Bassett, J., et al. (2012). Implementation of Xpert MTB/RIF for routine point-of-care diagnosis of tuberculosis at the primary care level. *South African Medical Journal, 102*(10), 805-807.

Engel, N., Davids, M., Blankvoort, N., Dheda, K., Pant Pai, N., & Pai, M. (2017). Making HIV testing work at the point of care in South Africa: a qualitative study of diagnostic practices. [journal article]. *BMC Health Services Research, 17*(1), 408.

Engel, N., Davids, M., Blankvoort, N., Pai, N. P., Dheda, K., & Pai, M. (2015). Compounding diagnostic delays: a qualitative study of point-of-care testing in South Africa. *Tropical medicine & international health, 20*(4), 493-500.

Engel, N., Yellappa, V., Davids, M., Dheda, K., Pai, N. P., & Pai, M. (2018). Barriers to Point of Care Testing in India and South Africa. In S. Hostettler, S. Najih Besson & J.-C. Bolay (Eds.), *Technologies for Development: From Innovation to Social Impact* (pp. 75-85). Cham: Springer International Publishing.

England, K., Masini, T., & Fajardo, E. (2019). Detecting tuberculosis: rapid tools but slow progress. *Public Health Action, 9*(3), 80-83.

Gidado, M., Nwokoye, N., Nwadike, P., Ajiboye, P., Eneogu, R., Useni, S., et al. (2018). Unsuccessful Xpert(®) MTB/RIF results: the Nigerian experience. *Public health action, 8*(1), 2-6.

Hanrahan, C. F., Clouse, K., Bassett, J., Mutunga, L., Selibas, K., Stevens, W., et al. (2015). The patient impact of point-of-care vs. laboratory placement of Xpert((R)) MTB/RIF. *INT J TUBERC LUNG DIS, 19*(7), 811-816.

McDowell, A., & Pai, M. (2016). Treatment as diagnosis and diagnosis as treatment: empirical management of presumptive tuberculosis in India. *Int J Tuberc Lung Dis, 20*(4), 536-543.

McDowell, A., Raizada, N., Khaparde, S. D., Rao, R., Sarin, S., Kalra, A., et al. (2018). "Before Xpert I only had my expertise": A qualitative study on the utilization and effects of Xpert technology among pediatricians in 4 Indian cities. *PLOS ONE, 13*(3), e0193656.

Miller, R., Parkhurst, J. O., Peckham, S., & Singh, R. B. (2012). A qualitative study of HIV testing and referral practices of private hospital doctors treating patients with TB in Chennai, India. *Int J Health Plann Manage, 27*(2), 180-191.

Mwaura, M., Kao, K., Wambugu, J., Trollip, A., Sikhondze, W., Omesa, E., et al. (under review). Situating trade-offs: Stakeholder perspectives on overtreatment versus missed diagnosis in transition to Xpert MTB/RIF Ultra in Kenya and Swaziland. *PLoS One*.

Pai, M., & Furin, J. (2017). Tuberculosis innovations mean little if they cannot save lives. *Elife, 6*.

Squire, S. B., Belaye, A. K., Kashoti, A., Salaniponi, F. M., Mundy, C. J., Theobald, S., et al. (2005). 'Lost' smear-positive pulmonary tuberculosis cases: where are they and why did we lose them? *International Journal of Tuberculosis and Lung Disease, 9*(1), 25-31.

Yellappa, V., Devadasan, N., Krumeich, A., Pant Pai, N., Vadnais, C., Pai, M., et al. (2017). How patients navigate the diagnostic ecosystem in a fragmented health system: a qualitative study from India. *Global Health Action, 10*(1), 1350452.

Final Report for WHO

Web Annex D.7. Diagnostic Accuracy of the Molbio Truenat Tuberculosis and Rifampicin-resistance Assays in the Intended Setting of Use

30 March 2020

www.ClinicaTrials.gov : NCT03712709

FIND
Campus Biotech
Chemin des Mines 9
1202 Geneva
Switzerland
T: +41 (0)22 710 05 90
F: +41 (0)22 710 05 99

CONFIDENTIALITY STATEMENT

CONFLICT OF INTEREST STATEMENT

FIND is a global non-profit organization that drives innovation in the development and delivery of diagnostics to combat major diseases affecting the world's poorest populations. Our work bridges R&D to access, overcoming scientific barriers to technology development; generating evidence for regulators and policy-makers; addressing market failures; and enabling accelerated uptake and access to diagnostics in low- and middle-income countries. A WHO Collaborating Centre, FIND works with more than 200 academic, industry, governmental, and civil society partners worldwide on projects that cross six priority disease areas.

All industry partnerships (including Molbio) are subject to review by FIND's Scientific Advisory Committee, or another independent review body; selection criteria for technologies and partnerships include due diligence, target product profiles and public sector requirements.

FIND also supports the evaluation of publicly-prioritized tuberculosis (TB) assays and the implementation of WHO-approved (guidance & PQ) assays. In order to carry out these evaluations, FIND has product evaluation agreements with several private sector companies, which strictly define its independence and neutrality vis-a-vis the companies whose products are evaluated, and which clearly define roles and responsibilities.
More information on our policy and guidelines for working with private sector partners can be found here: https://www.finddx.org/wp-content/uploads/2019/03/Private-Sector-Partners-Policy_PL-02-08-01_V1.1_Nov2018.pdf

Sponsor, investigators and partner institutions

Sponsor

Foundation for Innovative New Diagnostics (FIND)
Campus Biotech, Chemin des Mines 9, 1202 Geneva, Switzerland
Indian Council of Medical Research (ICMR) – National Institute for Research in Tuberculosis – India; No. 1 Mayor Sathiyamoorthy Rd, Chetpet, Chennai, India

Partner institutions

FIND, Geneva, Switzerland (organizing institution)
FIND-India, New Delhi, India (organizing institution)
Indian Council of Medical Research – National Institute for Research in Tuberculosis, Chennai, India (organizing institution)
Burnet Institute, Melbourne, Australia (organizing institution for Papua New Guinea site)
PD Hinduja Hospital and Medical Research Centre, Mumbai, India (FIND-coordinated study site)
State TB and Demonstration Center, Civil Hospital Campus, Asarwa, Ahmedabad, Gujarat, India (FIND-coordinated study site)
Intermediate Reference Laboratory, Guwahati Medical College, Bhangagarh, Guwahati, Assam, India (FIND-coordinated study site)
Universidad Peruana Cayetano Heredia, Lima, Peru (FIND-coordinated study site)
Ethiopian Public Health Institute, Addis Ababa, Ethiopia (FIND-coordinated study site)
Central Public Health Laboratory, Port Moresby General Hospital, Papua New Guinea (Burnet-coordinated study site)
Ospedale San Raffaele, Milan, Italy (discordance analysis study)
Rutgers New Jersey Medical School, Newark, New Jersey, USA (analytical study laboratory)
Molbio, Bangalore, India (manufacturer)

For quality control purposes the manufacturer had access to their own data that was generated in the course of the study, however, the manufacturer was not involved in the study plan, data analysis or interpretation of the results.

Section 2: Clinical Evaluation Study

Principal Investigators (listed alphabetically)

Principal Investigator – Melbourne, Australia
Dr Philipp du Cros
Burnet Institute,

Principal Investigator – Port Moresby, Papua New Guinea
Dr Evelyn Lavu
Central Public Health Laboratory, Port Moresby General Hospital

Principal Investigator – Chennai, India
Dr N.S. Gomathi
National Institute of Research in Tuberculosis

Principal Investigator – Lima, Peru
Dr Eduado Gotuzzu
Universidad Peruana Cayetano Heredia

Principal Investigator – Addis Ababa, Ethiopia
Abyot Meaza
Ethiopian Public Health Institute

Principal Investigator – Mumbai, India
Dr Camilla Rodrigues
P.D. Hinduja National Hospital and Medical Research Centre

Co-Investigators

Collaborator – Guwahati, Assam, India
Dr N.J. Das
Directorate of Health Services, Assam

Collaborator – Ahmedabad, Gujarat, India
Dr Pranav Patel
State TB and Demonstration Center, , Asarwa, Ahmedabad, 380016, Gujarat

Collaborator – Lima, Peru
Dr César Augusto Ugarte Gil
Universidad Peruana Cayetano Heredia

Collaborator – Chennai, Tamil Nadu, India
Dr Lavanya
District TB Centre, Chennai

Section 3: Ease of use study

Dr Paresh Dave
State TB Officer,
Gujarat, India

Dr N.J. Das
State TB Officer
Assam, India

Dr N.S. Gomathi
National Institute of Research in Tuberculosis
Chennai, India

Dr J. Lavanya
District TB Officer
Chennai, Tamil Nadu

Dr Yogesh Singh
Director Research and Development, Head, Molecular Diagnostics, Dr. B. Lal Clinical Laboratory, Jaipur, India

Acknowledgements

The development of this document was led by FIND, authored by Adam Penn-Nicholson and Samuel Schumacher, with contributions from Aurélien Mace, Pamela Nabeta, Morten Ruhwald, Claudia Denkinger, Sophia Georghiou, Stefano Ongarello, Sanjay Sarin, Sarabijt Chadha, N.S. Gomathi, Manjula Singh, Daniela Cirillo and David Alland. We would like to thank all the study participants and their families for generously volunteering to participate in this study. In addition, we would like to thank the study sites for their time and effort in conducting the study, and assisting with the analysis of the operational data. We are also grateful to The Bill & Melinda Gates Foundation, the Indian Council for Medical Research, and the India TB Research Consortium for providing funding for the study.

Abbreviations

BCG	Bacillus Calmette-Guérin
CFU/mL	Colony-forming unit/milliliter
CI	Confidence interval
CRF	Case report form
Ct	Cycle threshold
DMC	Designated microscopy centre
DNA	Deoxyribonucleic acid
DR	Drug resistant
DST	Drug-susceptibility testing
Dx	Diagnosis
FIND	Foundation for Innovative New Diagnostics
INH	Isoniazid
IVD	In-vitro diagnostic
LOD	Limit of detection
LPA	Line probe assay
LIMS	Laboratory information management system
LMICs	Low- and middle-income countries
MC	Microscopy Centre
MDR-TB	Multidrug resistant tuberculosis
MGIT	Mycobacterial Growth Indicator Tube
MTB	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* complex
NC	Negative control
NGS	Next generation sequencing
NTM	Nontuberculous mycobacteria
PBST	Phosphate buffered saline with Tween-20
PC	Positive control
PCR	Polymerase chain reaction
PQ	Pre-qualification
QC	Quality control
RIF	Rifampicin
SOP	Standard operating procedure
TB	Tuberculosis
TPP	Target product profile
WGS	Whole genome sequencing
WHO	World Health Organization
XDR	Extensively resistant tuberculosis

EXECUTIVE SUMMARY

Background

Xpert® MTB/RIF ('Xpert') has revolutionized the diagnosis of both tuberculosis (TB) and resistance to rifampicin (RIF) and the Xpert® MTB/RIF Ultra ('Ultra') was developed to achieve even higher sensitivity. However, these tests are run on the GeneXpert instrument, which requires operation in a temperature-controlled environment and is susceptible to dust (1). Given sustained, high rates of pre-treatment loss to follow-up (LTFU) (2), bringing sensitive TB diagnosis closer to patients is a key priority for global TB control (WHO High-priority TPPs (3)). This requires robust point-of-care diagnostic tests that are easily implementable at lower levels of the health care system.

Molbio Diagnostics Pvt. Ltd. (Bangalore, India) developed two assays that utilize chip-based real-time micro PCR for detection of TB and one assay for the detection of RIF resistance: the Truenat™ MTB (including both the MTB (*nrdZ* target) and MTB plus (*nrdZ* and *IS6110* targets)) assays for TB detection and the MTB-RIF Dx reflex assay for detection of RIF resistance. All three assays are run on DNA eluate, obtained from the automated Trueprep DNA extraction device that uses a universal cartridge-based system to extract DNA from 0.5mL of sputum in under 20 minutes. The DNA eluate is loaded onto the chip-based Truelab micro PCR device to detect the presence of *Mycobacterium tuberculosis* (MTB) DNA in the participant specimen in approximately 40 minutes. If MTB is detected, the Truenat MTB-RIF Dx reflex test can be run in the Truelab machine using the same DNA eluate. Both the Trueprep and Truelab devices are portable, battery-operated and can function within a wide range of environmental conditions.

Herein we report updated results of a multi-centre diagnostic accuracy study of the Truenat MTB, Truenat MTB Plus and MTB-RIF Dx assays performed at the microscopy centre level, alongside results from an assessment of the operational characteristics of the Truenat assays. While enrolment has been completed, this report is based on data from all returned culture results; final results will include 97 additional culture results from the 181 total participants enrolled at the site in Papua New Guinea, which were not available at the time of this analysis.

The report contains four sections: SECTION 1 provides a background to the Truenat systems; SECTION 2 reports on a multicentre prospective clinical evaluation; and SECTION 3 on operational characteristics.

Methods

A multicentre prospective clinical evaluation study was conducted in 19 clinical sites (each with a microscopy centres attached) and 7 reference laboratories in 4 countries to determine the diagnostic accuracy of the Truenat assays when performed in the intended settings of use (i.e. microscopy centres), relative to microbiological confirmation (culture) as the reference standard. The performance of the Truenat assays was also compared head-to-head (on the same specimens) to Xpert or Ultra in reference laboratories as part of this assessment. All sites performed Xpert, apart from sites in Peru, which performed Ultra. The analysis, presented here, reports on the results for 1,654 eligible participants with complete data (out of 1,925 participants who completed enrolment). Analysis of the full dataset will be written and submitted in 2020.

This report also describes an assessment of the operational characteristics, ease of use associated with the Truenat assays. For this assessment, data on the operational characteristics were provided by the manufacturer as well as collected from operators at study sites through questionnaires.

Results

Participants enrolled in this study were adults presenting to clinics with symptoms suggestive of TB disease, either without any prior treatment for TB in the last 60 days (Case Detection Group), or having received but not responding to treatment (Drug-resistant Risk Group). At the time of this analysis, enrolment had been completed and culture results were available for 1,654 participants, which form the basis of all reported analyses on diagnostic accuracy. The proportion of participants testing culture-positive across all sites was 24% (n=393), with 16% (n=62) of TB patients being RIF-resistant.

Overall, for sputum tested in microscopy centres, sensitivity of the Truenat MTB assay was 73% (95%CI 68, 78) and sensitivity of Truenat MTB Plus assay was 80% (95%CI 75, 84); among smear-negative specimens, sensitivities were 37% (95%CI 27, 48) and 46% (95%CI 36, 57), respectively. The specificities of Truenat MTB and MTB Plus was 98% (95%CI 97, 99), and 97% (95%CI 95, 97), respectively. The Truenat MTB Plus assay showed higher sensitivity than the Truenat MTB (sensitivity difference = +6.5 [95%CI +3.3, +10.7]), with somewhat lower specificity (specificity difference = -1.4% [95%CI -2.6, -0.4]).

The total error rate for the Trueprep DNA extraction system was 2.4%; upon retesting 87.5% of non-determinate results resolved. The proportion of non-determinate Truenat MTB and MTB Plus assay results on the initial test was 6.2% and 9.2%, respectively; this resulted in 6.2% of participants having no valid Truenat MTB results and 11.8% having no valid MTB Plus assay result on specimens tested in the microscopy centre on initial test. When allowing for a single repeat-test in the microscopy centre, 1.7% and 3.9% of participants remained with non-determinate results for Truenat MTB and MTB Plus, respectively. The non-determinate rates for Xpert and Ultra on initial testing were 2.6% and 0.0%, respectively.

In raw sputa tested in reference laboratories and split for comparative testing, the sensitivities of Truenat MTB, Xpert, Truenat MTB Plus, and Ultra were 84%, 85%, 87% and 96%, respectively; specificities were 97%, 97%, 95% and 97%, respectively.

The sensitivities of Truenat MTB-RIF Dx and Xpert assays for RIF-resistance detection were 82% (95%CI 67, 91) and 84% (95%CI 67, 93), respectively; and specificity was 98% (95%CI 94, 99) for both assays. In Peru (the only site where Ultra was used) sensitivity was 100% (95%CI 68,100) and specificity 96% (95%CI 86, 99) for both Truenat MTB-RIF Dx and Ultra tests. The non-determinate rate of Truenat MTB-RIF Dx varied greatly depending on whether reflex testing was done based on positive results from the Truenat MTB assay (6.7% non-determinate) or the more sensitive MTB Plus assay (15% non-determinate).

Specificity of both the Truenat and Xpert assays was reduced at comparable levels in individuals who presented with a prior history of TB disease. As only 10 HIV-infected participants with active TB had been enrolled for these analyses, the sample size was too small to evaluate the effects of HIV co-infection.

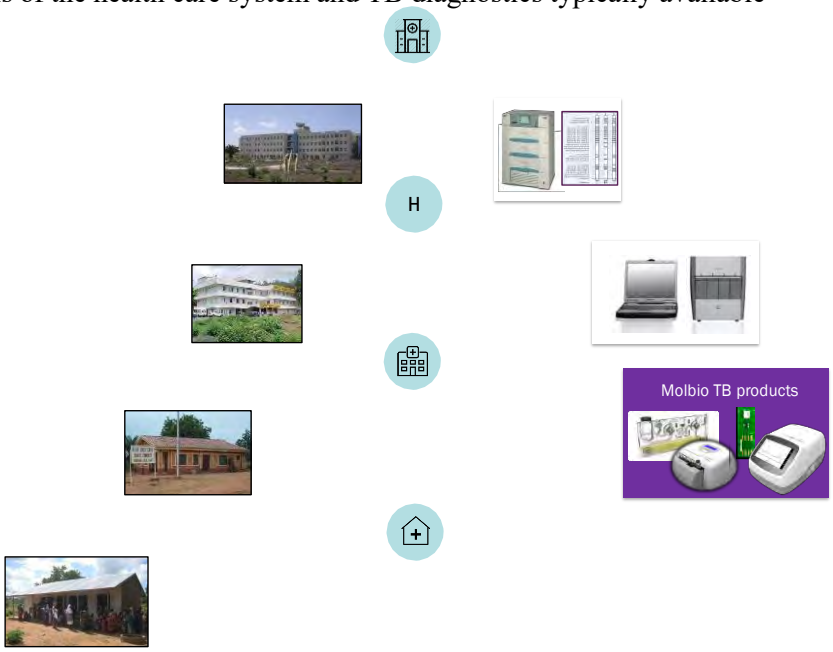
Extensive testing for quality control purposes was done. The overall proportion of positive test results for daily negative controls and weekly swab tests was 0.85% (17 of 1,993 tests) and 1.43% (14 of 963 tests), respectively. All such positive results were resolved upon cleaning and retesting, and the timing of positive swab and control results did not correspond with days where false-positive Truenat results on clinical specimens occurred.
User feedback from the study of operational characteristics was positive, with most users reporting appropriate setup and training, ease of use and ability to follow the manufacturers troubleshooting instructions for the assays. Nevertheless, some concerns were raised regarding the required testing time and effort, especially in high workload settings and when only a single-chip Truelab instrument was used.
Conclusion
For MTB detection, this analysis of the prospective clinical and laboratory validation studies suggests that the Truenat MTB and MTB Plus assays may have accuracy in a similar range to the Xpert and Ultra and can be performed in microscopy centres. Non-determinate rates were higher than for Xpert and Ultra, varying widely between sites. The results from the Truenat MTB-RIF Dx assay need to be interpreted with caution as data were limited. The Molbio platforms and assays have the potential to meet the minimal criteria set by the WHO TPP for a smear-replacement test.

SECTION 1: Introduction

Background

Rapid diagnosis and initiation of appropriate treatment is necessary to curb the spread of the TB epidemic. However, it is estimated that of the 10 million new TB cases in 2018, up to 3.0 million cases went undiagnosed (4) and the emergence of multi- and extensively drug-resistant TB (M/XDR-TB) has further complicated TB control efforts. Conventional culture and drug susceptibility testing (DST) methods rely on the slow growth of *Mycobacterium tuberculosis* (MTB) in solid or liquid media, which take weeks to months to yield results and can lead to prolonged periods of ineffective therapy and ongoing disease transmission. Furthermore, many countries with high TB burdens lack the resources to establish the stringent laboratory conditions needed for these growth-based methods and must rely upon smear microscopy tests which, at best, detect only 45% of TB infections (5). In 2018, approximately 484,000 people were diagnosed with rifampicin-resistant TB worldwide, of which 78% had MDR-TB (6). However, only 51% of all new TB cases diagnosed in 2018 were tested for resistance to rifampicin (RIF), one of the most important first-line anti-TB drugs (6). In view of the increasing incidence of M/XDR-TB, the development of rapid molecular diagnostic tests for the identification of MTB and resistance to RIF at the microscopy centre level has become a development and implementation priority (Figure 1).

Figure 1: Levels of the health care system and TB diagnostics typically available



Note: Images are for illustrative purposes only. The intended setting of use for Molbio Truenat is the Microscopy Centre.

Description of index tests

This report focuses on the following Molbio devices and diagnostic tests (instructions for use provided at http://www.molbiodiagnostics.com/products-listing.php):

- Trueprep Auto DNA extraction system
- Truelab DuoDx and Truelab QuattroDx micro-PCR machines
- Truelab MTB chip
- Truelab MTB Plus chip
- Truelab MTB-RIF Dx chip

The Truenat MTB and MTB Plus assays and the RIF reflex assay (Truenat MTB-RIF Dx) (Molbio Diagnostics, India), use real-time micro PCR for detection of MTB and selected RIF resistance in DNA extracted from a patient's sputum specimen (7) (Figure 2). The assays use automated, battery-operated devices to extract, amplify, and confirm the presence of specific genomic DNA loci, allowing for the rapid diagnosis of TB infections with minimal user input. These products are intended to be operated in peripheral laboratories with minimal infrastructure and minimally trained technicians can easily perform these tests routinely in their facilities and report PCR results in less than an hour. Moreover, with these devices PCR testing can also be initiated in the field level, on site. If the MTB assay result is positive, the user may then take another aliquot of extracted DNA and run the RIF-Dx assay to detect the presence of selected RIF resistance-associated mutations. The diagnostic performance of these assays has been previously evaluated in microscopy centres in India (8,9), but a larger assessment of the operational characteristics and acceptability of the technology is needed in intended settings of use (microscopy centre level) to confirm assay performance.

Figure 2: Truenat MTB assay steps

A mixture of raw sputum and liquefaction buffer is directly loaded onto the Trueprep Auto chip interface, which extracts MTB DNA in 18 minutes. The extracted DNA is transferred to the Truenat MTB (or Truenat MTB Plus) chip and then onto the Truelab Dx PCR machine, which detects the presence of MTB DNA and provides an automated result of either MTB-detected, MTB-not detected or non-determinate. For MTB positive results, another aliquot of the same DNA extraction is then transferred (reflex) to the Truenat MTB-RIF Dx chip.

Purpose of this study
The purpose of this study is to generate prospective clinical evidence to inform a WHO expert review on the diagnostic accuracy of the point-of-care Truenat MTB assays (MTB and MTB Plus) and the RIF-resistance reflex assay (MTB-RIF Dx) for a microscopy level setting, using a culture reference standard and comparator of Xpert and Ultra. The referenced study was a prospective, multicentre, diagnostic accuracy study in which the performance of an investigational rapid molecular diagnostic test (index test) on sputum specimens (Truenat MTB assays and RIF assay) was assessed in four countries using solid and liquid culture as the reference standard for the diagnosis of TB, and MGIT SIRE as the reference standard for the detection of RIF resistance.
With view to a meeting convened by WHO, the primary focus of this report was the provision of data to address the following 7 PICO questions for the three Truenat assays:

PICO questions for Truenat MTB assay
1. Should Truenat MTB be used to diagnose pulmonary tuberculosis in adults with signs and symptoms of pulmonary TB, against a microbiological reference standard
2. Should Truenat MTB be used to diagnose pulmonary tuberculosis in smear-positive adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?
3. Should Truenat MTB be used to diagnose pulmonary tuberculosis in smear-negative culture-positive adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?

PICO questions for Truenat MTB Plus assay
4. Should Truenat MTB Plus be used to diagnose pulmonary tuberculosis in adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?
5. Should Truenat MTB Plus be used to diagnose pulmonary tuberculosis in smear-positive adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?
6. Should Truenat MTB Plus be used to diagnose pulmonary tuberculosis in smear-negative, culture-positive adults with signs and symptoms of pulmonary TB, against a microbiological reference standard?

PICO questions for Truenat MTB-RIF Dx assay
7. Should Truenat MTB-RIF Dx be used to diagnose rifampicin resistance in adults with signs and symptoms of pulmonary TB?

Other analyses and results are provided for completeness and as supplementary information.

Description of comparator test: Xpert and Ultra
The Xpert® MTB/RIF assay on the GeneXpert (Cepheid, Sunnyvale, California, USA) platform is an automated nucleic acid amplification test which rapidly detects TB and resistance to RIF in less than two hours (10). Both detection of MTBC and resistance to RIF is done through targeting the *rpoB* gene of MTB. For TB detection, Xpert has demonstrated to have a pooled sensitivity of 89% (98% in smear-positive and 67% in smear-negative) and specificity of 99%, respectively (11). For the detection of RIF resistance, Xpert has a pooled sensitivity of 95% and specificity of 98%, respectively (11). In 2010, WHO endorsed Xpert and strongly recommended the assay be used as the initial diagnostic test for individuals with suspected MDR-TB or HIV-associated TB (12). The recommendations were expanded in 2014 for use in all patients, including extra-pulmonary TB and paediatric TB, following new evidence supporting use in these subpopulations (13). Xpert can be performed as a single use test by the GeneXpert IV or GeneXpert XVI systems or in a centralized capacity by the GeneXpert Infinity systems (48 or 80 tests). Xpert was used as the comparator assay in the analytical evaluation because (i) it is the TB assay with the most clinical performance data available and (ii) because its clinical sensitivity has been judged to be sufficient for use in all patients in whom TB is suspected (i.e. not restricted to smear-positive patients) and therefore was considered the more suitable benchmark test (based on WHO Expert Group Meeting reports, 2014 and 2016). The Ultra assay was used in Peru due to availability of the assay at the time of study initiation.

Initial analytical studies predicted the LoD of Xpert for MTB detection directly from sputum to be 131 colony forming units (CFU)/mL(14). More recently, the LoD for Xpert was shown to be 112.6 CFU/mL for TB detection and 200 CFU/mL for RIF resistance detection (15).

Despite the excellent ease of use and clinical performance of the assay, Xpert has a number of limitations. Its sensitivity in HIV-positive patients is estimated to be approximately 10% lower than for HIV-negative patients (i.e. 79%) (11), and assay sensitivity is also limited in those with paucibacillary disease (including patients with extra-pulmonary disease, children and those with early presentation) (16–18). In addition, limitations regarding the detection of RIF resistance have also been observed, with decreased sensitivity observed in hetero-resistant specimens (19,20) as well as in false-positive resistant calls seen for paucibacillary specimens (21). These limitations have mostly been

addressed by Ultra (22), although both assays have limitations in settings with environmental stressors, such as the need for a constant power supply and susceptibility to dust or high temperatures (23).

SECTION 2: Multicentre, diagnostic test accuracy study

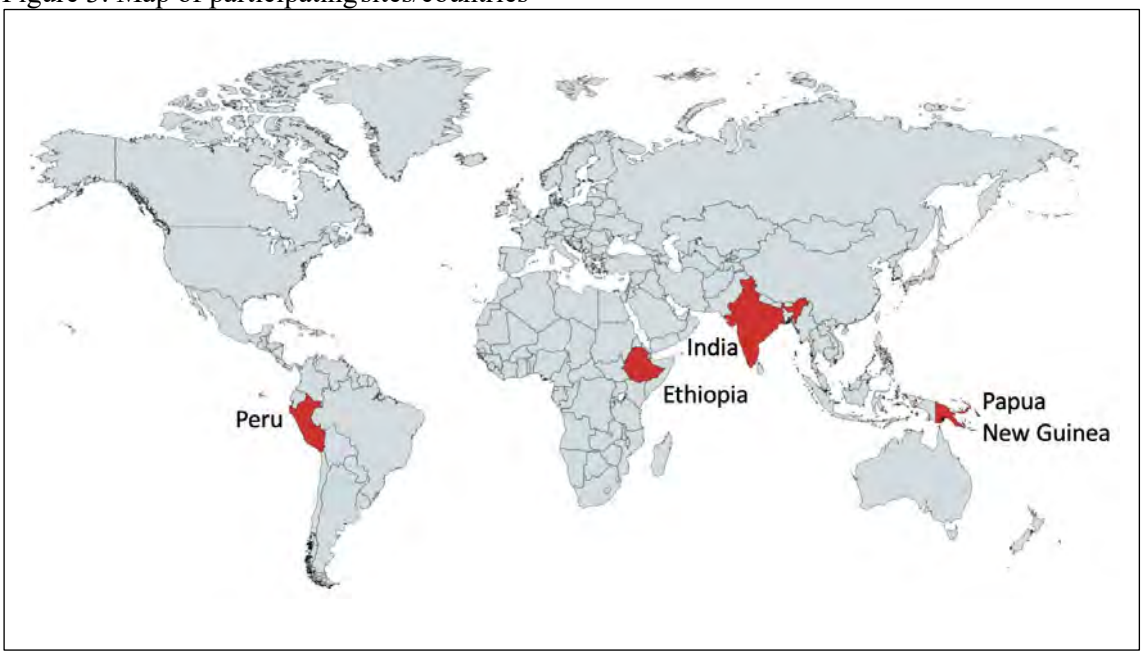
Methods
This was a prospective, multicentre, diagnostic accuracy study in which the performance of the Truenat MTB assays and MTB-RIF Dx assay – was assessed in four countries using solid and liquid culture as the reference standard for diagnosis of TB, and MGIT SIRE as the reference standard for the detection of RIF resistance.

Study objectives

Primary objectives
Estimate diagnostic accuracy of the Truenat assays (MTB and MTB Plus) for MTB detection among individuals undergoing evaluation for pulmonary TB, overall and per specimen, separately for smear-positive and smear-negative TB specimens, using a culture reference standard.
Estimate diagnostic accuracy of the Truenat MTB-RIF Dx assay for RIF resistance detection among individuals undergoing evaluation for pulmonary TB and DR-TB, using phenotypic DST as the reference standard.

Secondary objective
Compare the diagnostic accuracy of the Truenat assays (MTB and MTB Plus) and MTB-RIF Dx assay to that of Xpert, using a reference standard of culture for TB diagnosis and phenotypic DST for detection of RIF resistance.
Assess patient-important outcomes, including time to detection of TB and RIF resistance.
Study population and study sites
Study population: Men and women above 18 years of age presenting to clinics with symptoms suggestive of TB disease.
Study/sample size: The estimated enrolment need for the multicentre study was calculated to be at least 1,666 participants. The final enrolment target was of 1,882 participants.
Setting: This multicentre study was conducted in four countries and the enrolment planned was as follows:
India: 1,110 participants
9 clinics (with attached microscopy centres) and 1 private laboratory across 4 districts
Peru: 185 participants –> expanded to 400 participants
1 Reference lab and 5 clinics (with attached microscopy centres)
Ethiopia: 186 participants
1 Reference lab and 3 clinics (with attached microscopy centres)
Papua New Guinea (PNG): 186 participants
1 public hospital with clinic (with attached microscopy centre) and lab

Figure 3: Map of participating sites/countries



In India, a total of 9 clinics (with attached microscopy centres) were selected to represent the intended settings of use including urban, peri-urban/hilly, tribal and rural sites with low and high throughput laboratories. Given the important role of the private sector in India, 1 private laboratory (PD Hinduja Hospital, which is an DR-TB reference hospital) was included in the study.
The other three sites in South America, Africa and East Asia were selected to achieve wider geographic variation. These sites are outpatient TB clinics at district or regional health facilities (Figure 3 and Table 1).

Table 1: List of participating trial sites

India	Site 01	Mumbai: Hinduja
India	Site 02	Guwahati: Kamrup
	Site 03	Guwahati: Railway
	Site 04	Guwahati: Sonapur
	Ref Lab	Guwahati: Intermediate Reference Laboratory, Guwahati Medical College
India	Site 05	Chennai: Ayanavaram
	Site 06	Chennai: Villiwakkam
	Site 07	Chennai: Thanthai Perivar
	Ref Lab	Chennai: National Institute of Research in Tuberculosis
India	Site 08	Ahmedabad: Madhupura
	Site 09	Ahmedabad: CHC Chhala
	Site 10	Ahmedabad: PHC Kuha
	Ref Lab	Ahmedabad: Intermediate Reference Laboratory, State TB and Demonstration Center, Civil Hospital Campus
Peru	Site 11	Lima: CS Huascar II
	Site 12	Lima: CS Huascar XV
	Site 13	Lima: CS Jose Carlos Mariategui
	Site 14	Lima: CS Fraternidad
	Site 19	Lima: CS El Porvenir
	Ref Lab	Lima: Universidad Peruana Cayetano Heredia
Ethiopia	Site 15	Addis Ababa: Hiwot Amba
	Site 16	Addis Ababa: St. Gebrel
	Site 17	Addis Ababa: Woreda 01
	Ref Lab	Addis Ababa: Ethiopian Public Health Institute
Papua New Guinea	Site 18	Port Moresby: Central Public Health Laboratory, Port Moresby General Hospital

Participants were recruited sequentially at each clinic or through neighboring satellite clinic, and enrolled once informed consent was obtained, into one of two groups, namely a "Case Detection Group" and a "Drug-resistant TB Group". Sputum specimens were collected at the clinics and either sent to the centralized reference laboratory or processed and tested on site in the attached microscopy centres.
Trial participants met all of the inclusion criteria and none of the exclusion criteria (see Table 2).

Table 2: Inclusion and exclusion criteria

Case Detection Group	Drug-Resistant TB Group
Inclusion criteria	
Age 18 years or above Provision of informed consent Willingness to provide 3 sputum specimens (>2mL) at enrolment	
Willingness to have a study follow-up visit approximately 42 to 70 days after enrolment Clinical suspicion of pulmonary TB (including cough ≥2 weeks and at least 1 other symptom typical of TB)	Non-converting PTB cases (category I and category II failures) Retreatment cases* (those having failed a regimen, relapses or returned after loss to follow-up) Close contacts of DR-TB patients who have been diagnosed with active TB* Participants at high risk for MDR-TB as determined by local program*
Exclusion criteria	
Receipt of any dose of TB treatment within 60 days prior to enrolment Participants for whom, at the time of enrolment, the follow-up visit was poorly feasible (e.g. individuals planning to relocate)	Receipt of any MDR-TB treatment within 60 days prior to enrolment

* PTB cases on TB treatment are eligible if they are suspected to be treatment failures irrespective of how long TB treatment has been on-going. All culture-negative study participants on TB treatment were excluded from the analysis, even if they were smear-positive.

Additionally, participants who provided consent and who were enrolled, but did not provide a total of 3 sputum specimens (>2mL) were classified as early exclusions and withdrawn from the study.

Study procedures

For participants who were enrolled by the study team, the following information was recorded using standardized case report forms:

Demographic information

Targeted medical history, plus review of medical record, including (if performed for routine clinical care purposes) chest imaging results, CD4 T lymphocyte enumeration results, mycobacteriology laboratory results

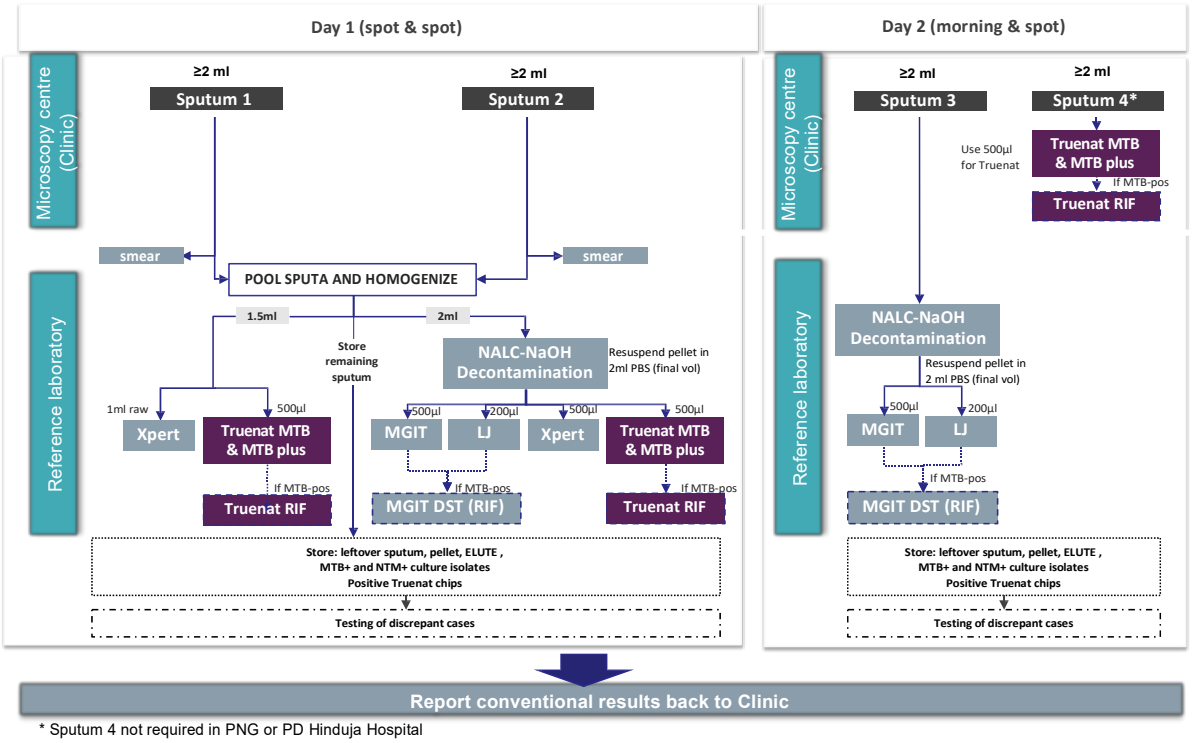
HIV test, unless any one or more of the following were available: written results of a positive HIV antibody test, written results of a positive HIV viral load, documentation in the medical record of positive HIV status by a treating clinician, immediate/verifiable documentation of HIV negativity within the preceding one month. HIV testing was performed using any test method approved by local health authorities following pre-test HIV counselling as per local guidelines

Participants were asked to provide four sputum specimens (S1, S2, S3, S4) over Days 1 and 2 (Figure 4). Each specimen had to be at least 2mL in volume. For the Case Detection Group, all specimens needed to be collected before the subject was started on TB treatment.

Note: as the sites at Hinduja Hospital in Mumbai, India, and the Port Moresby General Hospital in Papua New Guinea are centralized laboratory facilities, only 3 sputum specimens were collected as no microscopy centre was available

Laboratory testing was performed by index and reference standard tests as per specimen flow (Figure 4). Quality assured smear microscopy, culture and DST was performed on-site. GeneXpert systems for routine and study-specific Xpert testing was in place.

Figure 4: Specimen flow at enrolment



* Sputum 4 not required in PNG or PD Hinduja Hospital

Participants were enrolled at clinics/microscopy centres. All smears were read at the reference laboratories. Truenat testing occurred either in the reference laboratory (Day 1 sputa) or the microscopy centre (Day 2 sputa). Culture (liquid and solid) and subsequent drug sensitivity testing (DST) for rifampicin (RIF) was performed at the reference laboratories. Sputum4 was not collected at PD Hinduja Hospital or in Papua New Guinea (PNG). MGIT = Mycobacterial Growth Indicator Tube for liquid culture; LJ = Löwenstein Jensen solid culture.

On Day 1, each participant was asked to submit two spot sputa (S1 and S2, approximately 30-60min apart). Participants were given a labelled sputum pot and instructions for use, and asked to collect an additional sputum specimen (S3) the next morning (Day 2) before going to the clinic. At the clinic, participants were asked to provide a final spot sputum (S4). In the event that a participant failed to return on Day 2, S3 and S4 were permitted to be collected a maximum of 7 days after enrolment, provided that no TB treatment had been initiated (Case Detection Group).

Day 1: S1 and S2 – Two spot sputa were collected approximately 30-60min apart. A smear of each sputum specimen was prepared [17]. Thereafter, sputa totalling 4mL or more were pooled and homogenized by glass beads and vortexing in the reference labs. Homogenized sputa were further split: 1.5mL was used for analysis on raw/direct sputa, and at least 2mL used for NALC-NaOH decontamination.

Briefly, DNA was extracted independently from (i) raw sputum and (ii) decontaminated pellet by the Trueprep Auto device and tested on both the Truenat MTB and the MTB Plus chips, both of which were read by the Truelab real-time PCR 145nalyser. All DNA extracts testing positive by the MTB assay were subsequently tested by the Truenat MTB-RIF Dx assay (reflex), which was also read by Truelab 145nalyser.

Xpert assays were performed on the same raw and decontaminated specimens.

MGIT and LJ culture were performed only on the decontaminated specimen (Table 3).

Each positive culture was identified for MTB complex using MPT64 identification test and/or line probe assay (LPA). MGIT SIRE was used to determine the phenotypic DST for RIF.

Day 2: S3 – Morning sputum was returned to the clinic in a labelled sputum pot. S3 was sent to the reference laboratory and a second round of MGIT and LJ culture was performed on the decontaminated sediment.

Day 2: S4 – At the time that S3 was returned to the clinic, the participant was asked to provide spot sputum S4. The intended objective of this additional sputum specimen was to test the Truenat assay in the setting of use (i.e. microscopy centre). In both PD Hinduja Hospital, Mumbai and Central Public Health Laboratory (CPHL) Papua New Guinea, the microscopy centre and the reference lab are

the same. Thus, at these sites the Truenat assays were only performed once alongside Xpert (on Day 1).

Spot sputum S4 was processed in the microscopy centre: The entire volume of sputum was liquefied and lysed using Trueprep Auto kit reagents, and 500µL raw sputum was used for DNA extraction by Trueprep Auto and MTB detection by the Truenat assays. Any MTB-positive specimens were subsequently tested by the Truenat MTB-RIF Dx assay (reflex).
All positive Truenat chips were stored (refrigerated) at the FIND-coordinated sites to allow for sequencing from DNA amplicons if required for discordance resolution, as pre-defined in the protocol. Additionally, any leftover sputum, pellet, NTM+ or MTB+ culture isolates were stored (frozen).

Table 3: Reference standard test & index test procedures

Test	Notes*
Smear	Fluorescence microscopy (Auramine-O) or light microscopy (Ziehl Neelsen). Testing and reporting as per WHO/IUATLD guidelines (24).
Xpert	2:1 sample reagent added to raw sputum. In case of invalid, error or no result, testing was repeated if enough specimen was available
Ultra	2:1 sample reagent added to raw sputum and pellet (25). In case of invalid, error or no result, testing was repeated if enough specimen was available.
Liquid culture	Mycobacteria Growth Indicator Tube (MGIT) 960 culture; BD Microbiology Systems
Solid culture	Löwenstein Jensen. Testing and reporting done as per GLI mycobacteriology laboratory manual and local guidelines
MGIT DST	BD MGIT AST SIRE Test kit
LPA	Genotype MTBDR*plus*, Hain Lifescience, as per standard of care
MTB identification	MPT-64, SD Bioline, BD, or Capilia TB-Neo, TAUNS

*Testing done as per manufacturer's instructions unless otherwise specified

Follow-up & assessment of discordant cases

A follow-up visit at Day 56 (+/- 14 days) post-enrolment was conducted on a subset of participants in order to collect additional information on their TB status.
Culture-negative, Truenat MTB (and/or MTB plus) and Xpert/Ultra discordant cases During the prospective assessment, Truenat results were not provided to clinicians or participants or used for decision-making. Thus, participants who were Xpert-negative but Truenat-positive at enrolment ("discordant") were not treated on the basis of Truenat results. All culture negative cases with discrepant Truenat and Xpert/Ultra results underwent a follow-up visit performed at Day 56 (+/-14) post-enrolment for the following:
Interval medication history, including TB treatment and clinical evolution
An additional spot sputum specimen was obtained for smear microscopy and culture (LJ and MGIT) provided the participant had not been started on therapy and was able to provide a spontaneously produced sputum specimen
The intention was to aid in the identification of patients who would be diagnosed (in the absence of Truenat MTB assays being available for decision making) on clinical grounds.

Case Detection Group with negative results. The first 267 participants who were negative on all tests (approximately 20%) were scheduled for follow-up at Day 56 (+/-14) post-enrolment to assess:
Interval medication history, including TB treatment and clinical evolution
An additional spot sputum was obtained for smear microscopy and culture (LJ and MGIT) provided the pulmonary symptoms persist and the participant had not been started on therapy
The purpose of this follow-up visit was to identify the participants in this subset of those who were diagnosed or initiated on treatment on clinical grounds and those who were missed completely. A follow-up visit was not required for participants who were started on treatment (based on Xpert or culture results).
At the time of the writing of this report, no follow-up data was taken into account for analysis. Similarly, discordance analysis results are still pending further analysis using whole genome and targeted sequencing of DNA eluate, sediments, remaining sputa and PCR amplicons.
Analysis plan and statistical methods
Analysis datasets
Intention-to-test (ITT): all participants successfully enrolled in the study
Modified-Intention-To-Test (MITT) all participants in ITT for whom at least one test result is available
Per-Protocol (PP): all participants in ITT for whom results for all tests are available (complying with the protocol)
MTB Population (MTB_POP): all participants with uncontaminated culture results and without non-determinate test results, without any of the following:

no valid Truenat result for Truenat MTB and no valid result for Truenat MTB Plus
no valid culture result
2 contaminated cultures (unless other criteria for culture-positivity/negativity are met),
smear-positive, culture negative,
single positive culture with ≤20 colonies (LJ) or >28 days' time to positivity (MGIT)
culture-positive but no MTB complex identification available
specimens with growth of mycobacteria other than MTB complex only
RIF Population (RIF_POP): all participants with uncontaminated culture results and without non-determinate test results, without all of the above criteria for MTB_POP, and with:
For RIF detection, a valid phenotypic DST result for RIF

Table 4: Test status definition

Test result	Description
Smear-positive	≥ 1 positive smear (inclusive of scanty positive smears) using WHO grading
Culture-positive	≥ 1 LJ and/or MGIT culture growth confirmed MTB complex
Culture-negative	At least 2 LJ or MGIT have no culture growth after >56 days and >42 days
Contaminated culture	LJ: Cultures completely overgrown by bacterial or fungal contaminations within 3 weeks (discarded). In case of mixed cultures, isolated MTB colonies transferred to new LJ tube (repeat culture) MGIT: Instrument positivity without detection of AFB
Xpert-positive	MTB positive on Xpert® MTB/RIF
Xpert-negative	MTB negative on Xpert® MTB/RIF
Xpert-invalid	Any test run that is invalid, error, or inability to produce a result from a single Xpert® MTB/RIF run
Xpert RIF- indeterminate	MTB positive on Xpert® MTB/RIF with indeterminate for RIF-detection only
Xpert RIF-positive	MTB RIF-resistant result on Xpert MTB/RIF or Ultra assay
Xpert RIF-negative	MTB RIF-sensitive result on Xpert MTB/RIF or Ultra assay
Ultra-positive	MTB positive on Xpert® Ultra
Ultra-negative	MTB negative on Xpert® Ultra
Ultra-invalid	Any test run that is invalid, error, or inability to produce a result from a single Xpert® Ultra run
Ultra RIF-indeterminate	MTB positive on Xpert® Ultra with indeterminate for RIF-detection only
Truenat MTB-positive	MTB positive on Truenat MTB chip
Truenat MTB-negative	MTB negative on Truenat MTB chip
Truenat MTB-non-determinate	Any test run that is invalid, indeterminate, error, or inability to produce a result from a single Truenat MTB chip
Truenat MTB Plus-positive	MTB positive on Truenat MTB Plus chip
Truenat MTB Plus-negative	MTB negative on Truenat MTB Plus chip
Truenat MTB Plus-non-determinate	Any test run that is invalid, indeterminate, error, or inability to produce a result from a single Truenat MTB Plus chip
Truenat MTB-RIF Dx-positive	MTB RIF-resistant result on Truenat MTB-RIF Dx chip
Truenat MTB-RIF Dx-negative	MTB RIF-sensitive result on Truenat MTB-RIF Dx chip
Truenat MTB-RIF Dx non-determinate	Any test run that is invalid, indeterminate, error, or inability to produce a result from a single Truenat MTB-RIF Dx chip

Exclusion criteria for MTB and RIF detection analyses
Participants data set with any of the following criteria were excluded from the primary analyses of diagnostic test accuracy:
no valid Truenat assay result
no valid Xpert or Ultra result
no valid culture result

no valid phenotypic DST result for RIF (for RIF analysis only)
2 contaminated cultures unless other criteria for culture-positivity/negativity are met
smear-positive, culture-negative
single positive culture with ≤20 colonies (LJ) or >28 days' time to positivity (MGIT)
culture-positive but no MTB speciation available
specimens with growth of mycobacteria other than MTB complex only

Reference standards and case definitions (per-participant basis) for MTB and RIF
The reference standard for TB classification is based on TB culture and speciation results: a specimen is defined as TB positive if at least one of the culture results is positive and speciation confirms MTBC; a specimen is defined as negative if no culture is positive for MTBC and at least two culture results are negative (i.e. a single negative culture result with all other cultures contaminated does not suffice).
In addition, the following case definitions will be used for the final analyses of MTB and RIF detection. For MTB detection, the main analyses will be based on the three defined TB categories based on microbiological tests; case definitions using clinical information will be used in sensitivity analyses. For RIF detection, the main analyses will be based on phenotypic test results; genotypic test results will be used for sensitivity analyses.

The case definitions used for the analyses of MTB and RIF detection are shown in Table 5.

Table 5: Case definitions

DIAGNOSIS	DESCRIPTION
Smear-positive, culture-positive pulmonary TB	Patient with ≥ 1 positive smear (inclusive of scanty positive smears) and any positive culture result as per definitions of test results
Smear-negative, culture-positive pulmonary TB	Patient with all negative smears and any positive culture result as per definitions of test results
Microbiologically non-TB case	Smear- and culture-negative case as per definitions of test results
Non-TB case	Smear-negative, Xpert-negative and culture-negative and not started on TB treatment on the basis of clinical criteria. For Truenat-positive/Culture-discordant cases, a follow-up with repeated clinical and bacteriological work-up will be required to exclude TB with the highest possible likelihood. Only if the bacteriological work-up remains negative, the participant is called Non-TB.
Clinical TB case	Any participant who tests smear-negative, Xpert-negative, culture-negative but is started on TB treatment on the basis of clinical criteria and possibly other diagnostic tests such as chest-X-ray.
NTM	Culture-positive with NTM on rapid speciation test AND no other culture positive for MTB
Phenotypic RIF-resistant	Culture-positive and growth for Rif in conventional DST testing.
Phenotypic RIF-sensitive	Culture-positive and no growth for Rif in conventional DST testing

Metrics: sensitivity, specificity and predictive values
Point estimates and 95% confidence intervals (based on Wilson's score method) for sensitivity and specificity were derived based on the following definitions:

Table 6: Reference standard classification

	Reference standard classification			
Case prediction		Positive	Negative	Total
	Predicted positive	a	b	(a + b)
	Predicted negative	c	d	(c + d)
	Total	(a + c)	(b + d)	(a + b + c + d)

a = True Positives,
b = False Positives
c = False Negatives
d = True Negatives

Sensitivity = a / (a + c)
Specificity = d / (b + d)

Analysis of the primary outcome(s)
The primary objectives were analysed with the methodology described above to determine Truenat diagnostic accuracy.

Primary objectives
MTB diagnostic accuracy
Estimates of sensitivity and specificity of the Truenat MTB assays were calculated on the MTB_POP population (both smear positive and smear negative), using as reference standard of ≥ 1 culture positive specimen. The analysis was also stratified by smear-status (smear-positive culture-positive pulmonary TB vs. smear-negative culture-positive pulmonary TB); by specimen (raw/direct sputum vs. sediment from the reference lab on Day 1); and by setting (reference lab vs. microscopy centre, comparing Day 1 reference lab sputum to Day 2 microscopy centre sputum).

RIF-resistance diagnostic accuracy
Estimates of sensitivity and specificity of the Truenat MTB-RIF Dx assay were calculated on the RIF_POP population, using results of the MGIT DST as the reference standard.

Analysis of the secondary outcome(s)
Secondary objectives
Estimates of sensitivity and specificity of the Truenat MTB assays and MTB-RIF Dx assay were calculated for the MTB_POP population (both smear positive and smear negative), based on the methodology described above; the results were then compared with Xpert and Ultra, using culture as the reference standard. Comparisons were determined based on identical specimen types, i.e. comparing Truenat assays and Xpert on direct sputum from Day 1, and similarly comparing Truenat assays and the Xpert assays on decontaminated sediment from Day 1, using culture from Day 1 or Day 2 as the reference standard.

In order to evaluate the difference in performance between the two tests, the difference of the proportions was also reported, together with a 95% confidence interval based on Tango's score method.

Outcome
We report on the number of TB (culture-positive) cases detected by the Truenat MTB assay, by smear (any positive grade) and by Xpert. We also report on the difference in proportion of Truenat detection vs. the other methods.

We report on the number of RIF-resistant cases detected using the Truenat RIF assay, together with the number of RIF-resistant cases detected following the current standard of care. This allows for an estimation of number of RIF-resistant cases that are missed by using smear and Xpert vs. the Truenat test.

Sample size and enrolment targets
The sample size was set with the aim to achieve high confidence in the accuracy estimates for MTB-detection and RIF resistance detection for the overall multi-country study.
Based on an expected sensitivity of 67% with Truenat MTB Plus for detection of TB among smear-negative/culture-positive cases (based on preliminary data), 80 smear-negative/culture-positive cases would be required to achieve a total width of the 95% confidence interval of 20% (95%CI: 57 to 77). Assuming a TB prevalence of 20% and a 30% prevalence of smear-negative/culture-positive TB cases, the total number of participants to be enrolled would be 1,333. To account for losses, this was inflated by 20%, yielding a final sample size of 1,666 participants under investigation for TB overall. Two thirds of the study participants were planned for recruitment in India, i.e. n = 1,110 thus 67 smear-negative/culture-positive cases and a 95% confidence interval of 21% could be achieved (95%CI 55 to 79).The other one-third of enrolled participants (n = 556) were planned for recruitment in three other countries in order to provide geographic variation.
A numerical simulation based on an expected sensitivity of Truenat MTB Plus and Xpert of 67% among smear-negative/culture-positive specimens, with a correlation of 0.5 between the two tests, indicates that with the planned sample size, a 95% confidence interval of +/- 11% should be expected on the estimate of the percentage difference in performance between the two tests.

The sample size for the secondary objective of determining diagnostic accuracy of RIF resistance by Truenat MTB was selected based upon an expected Truenat RIF sensitivity of 95% with a confidence interval of 10% (90-100%), requiring at least n = 37 RIF-resistant participants detected. We assumed a prevalence of 20% culture-positive TB cases detected across all presumed TB cases, 2.8% RIF resistance amongst all culture-positive TB cases, and 12% prevalence of RIF resistance amongst TB retreatment cases. We thus predicted that 1,542 re-treatment patients would need to be enrolled. While the prevalence of culture-positive TB cases might have been higher if enrolment were to be conducted at drug-resistance TB referral clinics, this would have compromised primary objectives for assessment in the intended settings of use. As such, it was possible that the sufficient sample size to allow for the analysis of secondary objectives in this study would not be reached. For this reason, detection of RIF resistance was continually monitored throughout the study. The identified shortfall is currently being supplemented with an analytical sub-study using confirmed genotypically and phenotypically well-characterised RIF-resistant strains from the FIND specimen bank. For the entire study there was an enrolment cap of 200 participants in the Drug Resistance Group in order to avoid undermining the primary objective of enrolling smear-negative culture positive TB cases.

Quality assurance

The open-nature of the Truenat assay system increases the potential for cross-contamination. Therefore, to better monitor this risk, all sites performed daily negative control and weekly swab testing of both the Truelab and Trueprep machines. These QC steps were run at the end of the day to detect any potential contamination after a day's work, opposed to any remaining contamination following the morning cleaning procedures. If any negative control or swab was identified as positive, a round of cleaning (with 0.5% sodium hypochlorite and removal with 70% ethanol) was conducted, followed by repeat testing using individual swabs.

External controls testing

External negative controls were tested throughout the study to assess the impact of DNA amplicon contamination on the performance of the Truenat system. One negative control was run for every day of Truenat testing and for each Truelab machine tested. The negative control consisted of the Trueprep lysis buffer reagent used without sputum, which was processed on the Trueprep DNA extraction device. The eluate was subsequently run on each Truelab micro PCR machine on a Truenat MTB Plus chip, alternating loading bays on the Truelab Duo-Dx or Quatro-Dx. Troubleshooting of unexpected results included additional cleaning steps and subsequent swab testing, as below.

Swab testing

In addition to the testing of external controls, swabs were tested on Truelab machines on a weekly basis and whenever external controls provided positive results, to identify potential DNA amplicon contamination of the working areas. Briefly, separate sterile swabs dipped in sterile water were used to specimen the Truelab processing area and system surfaces. The swabs were then "pooled" into a single tube containing Trueprep lysis buffer, and processed through the Trueprep and Truelab systems. As for the external controls testing procedures, troubleshooting of unexpected swab results included additional cleaning steps and individual swab sampling and testing.

Non-determinant assay analysis

The proportion of non-determinate results for the Truenat assay system was assessed in both clinics and reference laboratories. A non-determinate result was defined as any result that did not provide a valid result on either the Trueprep or Truelab equipment. These non-determinate results included both operator errors and equipment/software errors or failures, or invalid results or indeterminate results. Under any of these circumstances, the participant would not receive a valid result to definitely classify their sputum specimen as MTB detected or MTB not detected.

On the Trueprep DNA extraction device, a non-determinate result could include cartridge failure due to:

pressurization issues
valve failure
cartridge manufacturing fault
cartridge leakage
heating system failure
clogged cartridge due to sputum not being liquefied sufficiently
buffer loading shortage
A cartridge that was incorrectly loaded could be re-loaded and re-run, and thus would not constitute a non-determinate result.

For the Truelab equipment, a non-determinate finding may include:
An indeterminate result (no MTB detection call)
An invalid result (no detection of the internal positive control)
An equipment or operator error, which could be further stratified by:
Error 1: Thermal cycling or probe check error
Error 2: Test stopped manually by operator
Error 3: Error with optical profile
Error 4: Runtime error
Error 5: Invalid due to insufficient PCR enzyme.
Any chip that was incorrectly loaded onto the Truelab machine was replaced and rerun. Any chip where DNA was not loaded correctly onto the chip loading surface was discarded, replaced and reloaded with DNA. The results presented below do not capture errors in DNA loading or chip loading as site incident logs did not report high levels of such errors.

Discordant analysis

Given that the sensitivity and specificity of culture is not perfect, misclassification by the reference standard in this study may introduce bias into the accuracy estimates for the Truenat assays. Moreover, false-positive results for TB detection could result from cross-contamination, amplification of a wrong gene target or detection of dead bacilli by molecular tests.

The discordant analysis work plan for the study will commence after return of final culture results. Remaining DNA eluate, sputum, pellet and, where appropriate, amplicons from positive Truenat chips arising from participants with discordant Truenat and culture results will be tested by targeted sequencing. Targeted sequencing will be used to determine the true genotype of the specimen in question.

Additional measures, including swab testing of lab surfaces and testing of known negative specimens, will increase confidence in the validity of results obtained and to aid in the identification of potential sources of cross-contamination. The same number of non-discordant cases will also be tested to avoid reclassification bias.

Results based on initial testing only, as well as those obtained upon repeat testing, will be presented.

The full SOP for discordance analysis is provided in the Appendix.

Monitoring strategy

A combination of centralized, remote and on-site monitoring was conducted for this trial:
All sites had an initial on-site visit (site initiation visit), lasting at least 2 days and conducted by the FIND Trial Manager
All sites had a first interim/on-site visit within 6 weeks after the start of enrolment
All sites will have a close-out visit, either on site or remotely depending on available resources

Centralized monitoring is performed by FIND data management in Geneva

Data management

Data was captured through single data entry at the sites onto FIND's online clinical studys platform from paper-based case report forms (CRF). The system was password protected and data quality checks were performed on a regular basis to identify data that appear inconsistent, incomplete, or inaccurate. Quality control checks performed by Data Management included Edit and Range Checks programmed into OpenClinica.

FIND was ultimately responsible for compiling data and conducting the analysis. Key sections of statistical code were completely re-written and re-run by an independent statistician. Additionally, the entire statistical code was checked, analyses rerun, and results confirmed.

The full study protocols can be shared upon request.

Changes to the initial study procedures and analysis

Enrolment targets: The protocol initially intended to enrol 1,666 participants across 4 sites. However, due to slower than anticipated enrolment at some of the sites, and the lack of Truenat testing in microscopy centres for participants enrolled at Hinduja Hospital in Mumbai and in Port Moresby in Papua New Guinea, we chose to expand enrolment at two sites:
Lima, Peru = expanded from 185 to 400 participants
Chennai, India = overenrolled from 321 to 350 participants

The total revised target for enrolment is 1,882 participants will be terminated at all sites on 31 Dec 2019. As of 23 October 2019, 1,322 participants had been enrolled.

Secondary objectives: As a secondary objective, we had intended to report on patient-important outcomes such as the time from sputum collection to detection of TB in the microscopy centres using the Truenat MTB assays. However, in the context of this study where additional processing and procedures were required beyond that of routine standard of care, we identified that this analysis was unrealistic.

Results

Study population
Between March 2019 and Feb 2020, 1,925 participants consented to participate in the study, and 1,916 participants met the eligibility criteria for enrolment across the 19 study sites. At the time of this analysis, complete data (i.e. completed CRFs and sufficient sputa received for testing) was available for 1,700 participants. The main reason for exclusion from analysis for this report were incomplete CRFs, largely due to outstanding culture results or pending data entry. Forty-six participants were excluded as they were missing either culture results or all Truenat results, or had a smear-positive culture-negative result (possibly indicating over-decontamination of sediment); this was most relevant at the site in Guwahati, India, where 10 participants out of 265 eligible for analysis (3.8%) had smear-positive culture-negative results. In total, data for 1,654 participants were included for this analysis. Amongst these, 233 participants did not have a Day 2 sputum specimen collected for analysis of Truenat performance in the microscopy centre as they were enrolled at a site that did not have access to microscopy centres. In total, 246 participants were missing one or more valid Truenat test results (Figure 5).

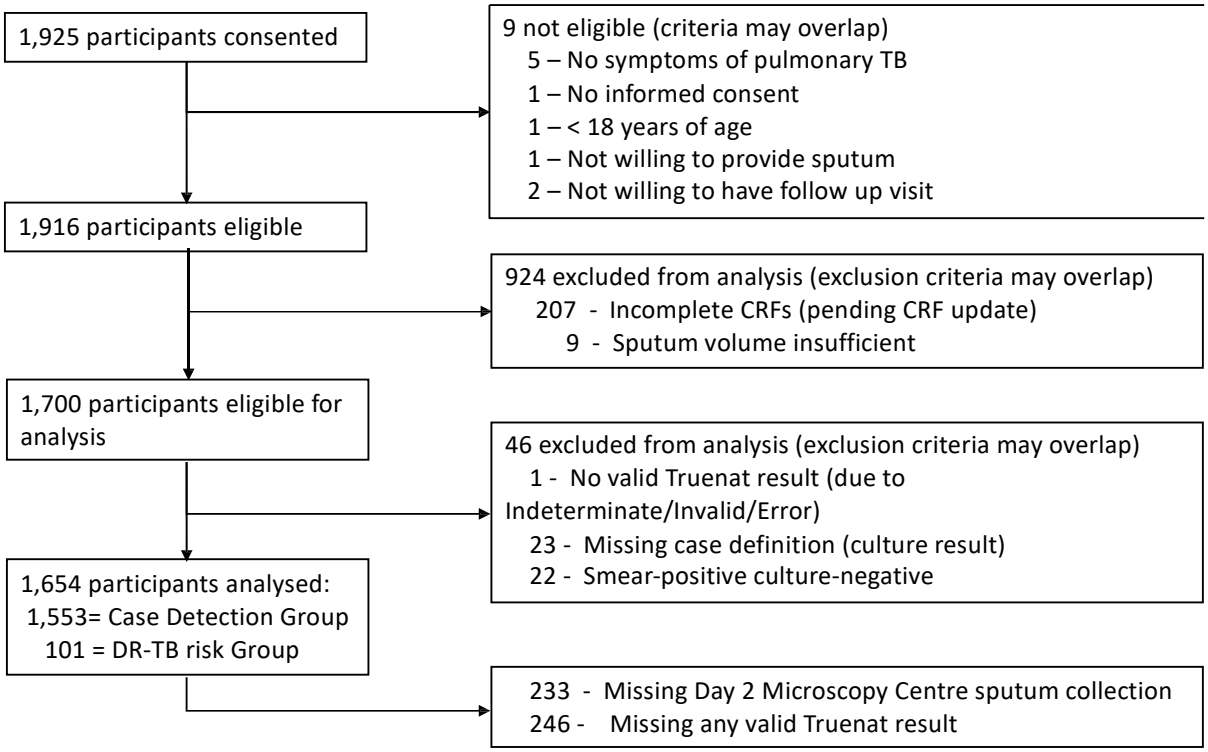


Figure 5: Participant exclusions flow chart
Note: Truenat non-determinate results are excluded from the accuracy analyses but are reported separately. Eligibility criteria stratified participants for inclusion in the Case Detection Group or the Drug Resistant TB (DR-TB) Risk Group as defined in the methods section.

Of the total 1,654 participants included in this analysis, 93%, (n=1,553) were enrolled into the Case Detection Group for analysis of accuracy for MTB detection. These participants, as well as the additional 101 participants enrolled into the DR-TB Risk Group (who were already on treatment regimens at the time of enrolment), were also eligible for the assessment of Truenat MTB-RIF Dx accuracy for RIF-resistance detection.

Demographic and clinical characteristics of the enrolled participant population are shown in Table 7 by site. The median age of participants was 40 years, with women making up 46% of the total participant population.

Table 7: Demographic and clinical characteristics of enrolled participant population

	All	HINDUJA	GUWAHATI	CHENNAI	AHMEDABAD	PERU	ETHIOPIA	PNG
N	1654	141	244	313	287	393	194	82
Age Med [min - max]	42 [18 - 88]	38 [18 - 86]	43 [18 - 82]	48 [19 - 83]	47 [19 - 85]	38 [19 - 88]	37 [18 - 81]	36 [19 - 78]
Female sex (%)	43.35	49.65	35.25	42.81	35.54	49.62	51.03	37.8
HIV-infected (%)	2.54	0.71	0	0	0.35	1.78	14.43	6.1
DR-TB Risk Group (%)	6.11	67.38	0	0	1.05	0.76	0	0
CulturePositive (%)	23.76	70.21	23.36	12.78	18.47	24.17	12.37	30.49
Smear positive (%)	16.75	56.74	17.21	7.99	15.33	13.74	8.25	19.51
Smear-positive Culture-positive (%)	69.97	78.79	73.68	62.5	83.02	56.84	66.67	64
Smear-negative Culture-positive (%)	30.03	21.21	26.32	37.5	16.98	43.16	33.33	36
Xpert Positive (%)	23.58	70.92	24.18	12.14	17.07	25.45	9.79	30.49
Xpert RIF Positive (%)	3.26	20.57	2.46	0.32	1.74	2.8	0.52	1.22
DST RIF Resistant among culture positive(%)	15.78	32.32	19.3	2.5	5.66	11.58	4.17	12

Note: Final culture results are yet to be returned from Papua New Guinea (PNG), as results were only available from 82 participants from 200 enrolled. TB positive is defined by either MGIT and/or LJ positivity with identification of MTBC.

HIV prevalence overall was 2.5%; in Ethiopia it was 14.4% and in PNG was 6.1%, although of those enrolled only 10 participants had active TB. The HIV-infection status was unknown for many participants but country-level reports of HIV co-infection prevalence among TB patients are 5% in Ethiopia, 6% in Peru, 3% in India and 7% in Papua New Guinea (4). The TB prevalence (based on the reference standard) was 24% overall, with 21% in the case detection group and 65% in the DR-TB risk group. TB prevalence was highest in Hinduja Hospital in Mumbai (70%, n=99/141) as this site is a TB referral centre and therefore also enrolled a larger percentage of participants into the DR-TB Risk Group.

Among the 327 culture-positive participants in the Case Detection Group, 33% tested negative by smear microscopy on both specimens tested.

The prevalence of RIF-resistance in culture-positive participants, based on phenotypic DST results, was 15.8% in total (14% among new cases and 24% among participants in the DR-TB risk group); as expected, there was large variation between sites, with highest rates of RIF-resistance (32%) noted in PD Hinduja Hospital, which is a DR-referral clinic.

The primary analyses focused on the diagnostic performance of the Truenat assays using culture results as the reference standard. MTB detection results were investigated for the Case Detection Group. Results for RIF detection were evaluated for all participants. Sensitivity results were stratified by smear status. Figure 6 graphically indicates the different cohorts, used for each analysis.

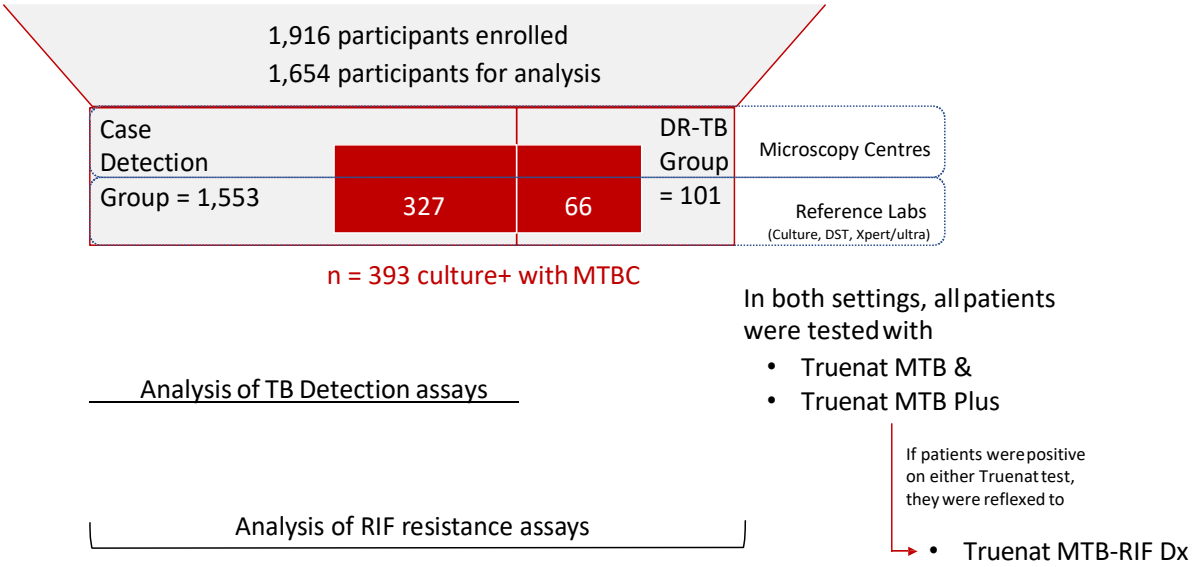


Figure 6: Schematic of Truenat performance analysis strategy.

A subset of 1,654 of the total 1,916 enrolled participants was included for this analysis of diagnostic accuracy of the Truenat assays. All enrolled participants provided sputum, which was tested by Truenat MTB and MTB Plus assays in both the microscopy centres and the reference laboratories. If either of these Truenat tests were positive, the specimen was reflexed to the Truenat MTB-RIF Dx chip for detection of RIF-resistance. Data from participants in the Case Detection Group (no TB treatment in the last 60 days) were used for analysis of the assays for TB detection (Truenat MTB and Truenat MTB Plus). Data from participants in both the Case Detection Group and the DR-TB Risk Group (having received first line TB treatment within 60 days of enrolment without improvement of symptoms) were used for analysis of RIF-resistance with the Truenat MTB-RIF Dx assay. To evaluate assay performance in the intended setting of use, the primary analyses focus on data from specimens processed in the microscopy centre. To compare Truenat assay performance to that of Xpert or Ultra, only specimens processed in the reference laboratories were used. Note: Truenat non-determinate results are excluded from the accuracy analyses but are reported separately. For the analysis, only participants with complete CRFs and culture results were included.

Diagnostic Accuracy of the Truenat MTB detection assays in the intended setting of use (results for PICO questions)

Of the 1,916 enrolled participants with complete data for this analysis, 1,553 formed part of the Case Detection Group (having not received any TB treatment in the 60 days preceding enrolment) and 1,336 of these participants had valid Truenat results for both the MTB and the MTB Plus assays processed at the microscopy centre, and had valid culture results with identification of MTBC. This group was used to assess sensitivity and specificity of the Truenat assays when used in the intended setting of use (microscopy centre) for MTB detection. Of these participants, 258 were culture positive with MTBC identification; 174 were smear-positive culture positive and 84 were smear-negative culture positive. For detection of RIF-resistance at the microscopy centre, all participants enrolled in both the Case Detection Group and the DR-TB Group (having received first-line TB treatment in the 60 days prior to enrolment, without symptom resolution) were included if they had valid Truenat and culture results. Of the participants with a positive Truenat MTB and MTB Plus assay, 260 were reflexed to Truenat MTB-RIF Dx test; 176 were smear-positive culture-positive and 84 were smear-negative culture-positive.

Sensitivity of microscopy centre sputum testing was 73.3% for Truenat MTB and at 79.8% for Truenat MTB Plus (Table 8 and Table 9). Specificity was 97.9% and 96.5% for Truenat MTB and MTB Plus, respectively. Sensitivity for smear-negative culture positive participant specimens was

36.9% for Truenat MTB and 46.4% for Truenat MTB Plus (Table 8). Contingency tables comparing Truenat MTB and MTB Plus are provided in the Appendix.

Table 8: Performance of Truenat assays for TB detection at the microscopy centre and for RIF-resistance detection at the microscopy centre and the reference laboratory

All participants	N	TP	FP	FN	TN	Sensitivity % (95% CI)	Sensitivity % Smear Pos (95% CI) - N	Sensitivity % Smear Neg (95% CI) - N	Specificity % (95% CI)
Microscopy Centre sputum									
Truenat MTB	1336	189	23	69	1055	73.3 [67.5,78.3]	90.8 [85.6,94.3] - N:174	36.9 [27.4,47.6] - N:84	97.9 [96.8,98.6]
Truenat MTB Plus	1336	206	38	52	1040	79.8 [74.5,84.3]	96 [91.9,98.0] - N:174	46.4 [36.1,57.0] - N:84	96.5 [95.2,97.4]
Truenat MTB Rif-Dx	186	16	8	3	159	84.2 [62.4,94.5]	87.5 [64.0,96.5] - N:16	66.7 [20.8,93.8] - N:3	95.2 [90.8,97.5]
Ref lab sputum									
Truenat MTB Rif-Dx	309	43	7	8	251	84.3 [72.0,91.8]	86.4 [73.3,93.6] - N:44	71.4 [35.9,91.8] - N:7	97.3 [94.5,98.7]

Note: Sensitivity/specificity for Truenat MTB and Truenat MTB Plus for TB detection (reference standard: MTB culture) and sensitivity/specificity for Truenat MTB-RIF Dx for RIF-resistance detection (reference standard: culture DST); Only participants in the Case Detection Group were included in the TB detection analyses for Truenat MTB and MTB Plus; whereas for RIF-detection analyses participants in both the Case Detection Group and the DR-TB Risk Group were included. Microscopy centre sputum was only available from participants enrolled at clinics with attached microscopy centres (n=1,553); among the 1,553 participants, 84 were smear-negative, culture-positive and 174 were smear-positive, culture-positive. Additional specimens were available for testing in the reference lab setting. Sensitivity of smear microscopy (based on two smears) was 68% (95%CI 63, 73). Note: Analysis of Truenat MTB-RIF Dx is shown on specimens collected at the microscopy centre and at the reference lab separately; most RIF-R participants were enrolled at PD Hinduja hospital which lacks a microscopy centre.

Comparison of the diagnostic accuracy of Truenat MTB and MTB Plus assays on the same sputum specimens in the microscopy centre showed higher sensitivity for Truenat MTB Plus assay than the Truenat MTB (sensitivity difference = +6.5% [95%CI +3.3, +10.7]), with lower specificity (specificity difference = -1.4 [95%CI -2.6, -0.4]) (Appendix).

Diagnostic Accuracy of the Truenat MTB rifampicin-resistance detection assay (results for PICO questions)

In the microscopy centre, the Truenat MTB-RIF Dx assay had 84.2% sensitivity and 95.2% specificity for RIF resistance detection (relative to RIF DST). However, these estimates are based on only three false-negative and eight false-positive results overall and given the limited sample size in this analysis, uncertainty around these estimates is high. Specimens collected at the DR-TB referral clinic at Hinduja Hospital in Mumbai contributed the bulk of RIF-resistant specimens in the study, but these did not contribute to the analysis above, which is focused on results from microscopy centres only. The MTB-RIF Dx assay done on sputum in the reference laboratories (i.e. the analysis including specimens from Hinduja Hospital) had a sensitivity of 84.3% (95%CI 72, 92) and specificity of 97.3% (95%CI 95, 99). For full assessment of diagnostic performance of the Truenat assays in the reference lab, please see the Appendix.

Additional analyses on performance of all Truenat assays at each site, for each specimen type, is included in the Appendix and comparative performance of the Truenat assays when tested on specimens directly or on NaOH decontaminated sediment are provided in the Appendix.

Diagnostic accuracy of Truenat assays compared to Xpert and Ultra

Next, we compared the performance of the Truenat assays to the Xpert and Ultra assays, using culture as the reference standard. As part of this study, all reference laboratories used Xpert as the comparator except the reference lab in Lima, Peru which used Ultra. Results show the performance of the Truenat assays relative to Xpert (Table 9 and Figure 7) and Ultra (Table 10 and Figure 8).

Table 9: Performance of Truenat assays for TB and RIF-resistance detection compared to Xpert

		N	TP	FP	FN	TN	Sensitivity % (95% CI)	Sensitivity % Smear Pos (95% CI) - N	Sensitivity % Smear Neg (95% CI) - N	Specificity % (95% CI)
Case Detection Group only	**Truenat MTB**									
	Xpert	1077	191	25	33	828	85.3 [80.0,89.3]	98.8 [95.7,99.7] - N:164	48.3 [36.2,60.7] - N:60	97.1 [95.7,98.0]
	Truenat MTB	1077	187	22	37	831	83.5 [78.1,87.8]	97.6 [93.9,99.1] - N:164	45 [33.1,57.5] - N:60	97.4 [96.1,98.3]
	Difference (Truenat MTB- Xpert)						-1.8 [-5.9,+2.1]	-1.2 [-4.7,+1.7]	-3.3 [-16.3,+9.5]	0.3 [-0.7,+1.4]
	Truenat MTB Plus									
	Xpert	1077	191	25	33	828	85.3 [80.0,89.3]	98.8 [95.7,99.7] - N:164	48.3 [36.2,60.7] - N:60	97.1 [95.7,98.0]
	Truenat Plus MTB	1077	195	39	29	814	87.1 [82.0,90.8]	98.8 [95.7,99.7] - N:164	55 [42.5,66.9] - N:60	95.4 [93.8,96.6]
	Difference (Truenat Plus MTB - Xpert)						1.8 [-1.9,+5.7]	0 [-2.8,+2.8]	6.7 [-6.0,+19.6]	-1.7 [-3.1,-0.3]
Case Detection and DR-Risk Group	**RIF detection**									
	Xpert Rif	218	32	4	6	176	84.2 [69.6,92.6]	88.6 [74.1,95.5] - N:35	33.3 [6.2,79.2] - N:3	97.8 [94.4,99.1]
	Truenat Rif	218	31	4	7	176	81.6 [66.6,90.8]	85.7 [70.6,93.7] - N:35	33.3 [6.2,79.2] - N:3	97.8 [94.4,99.1]
	Difference (Truenat Rif - Xpert Rif)						-2.6 [-13.5,+6.8]	-2.9 [-14.5,+7.3]	0 [-56.1,+56.1]	0 [-2.1,+2.1]

Note: Sensitivity/specificity for Truenat MTB and Truenat MTB Plus for TB detection (reference standard: MTB culture) and sensitivity/specificity for Truenat MTB-RIF Dx for RIF-resistance detection (reference standard: RIF DST). Only participants from the Case Detection Group were included in the TB detection analyses for Truenat MTB and MTB Plus (n=1,077,); whereas for the RIF resistance detection analyses participants in both the Case-detection Group and the DR-TB Risk Group were included (n=218). Data are shown for the reference lab sputum specimens only where comparative testing was done. Truenat performance comparisons were drawn against Xpert for participants from all sites except those enrolled in Peru where Ultra testing was conducted; Among the 1,077 participants with TB included for MTB detection analysis, 60 were smear-negative and culture-positive and 164 were smear-positive culture-positive.

In the Case Detection Group, comparison of Xpert to Truenat MTB and Truenat MTB Plus was available for 1,077 participant specimens, based on valid Truenat, Xpert and culture results. Performance of Truenat MTB and Truenat MTB Plus was largely comparable to that of Xpert. The Truenat MTB assay had marginally lower sensitivity for MTB detection than Xpert (difference = -1.8% [95%CI -5.9, +2.1]) but uncertainty around this estimate was relatively high. The Truenat MTB Plus assay had somewhat higher sensitivity than Xpert (difference = +1.8% [95%CI -1.9, +5.7]) with similarly high uncertainty (Table 9 and Figure 7). Specificity of Truenat MTB was similar to that of Xpert, while Truenat MTB Plus specificity appeared marginally lower than that for Xpert (difference = -1.7 [95%CI -3.1, -0.3]).

For the detection of RIF-resistance, DNA from 218 Truenat MTB- or MTB Plus-positive results from the DR-TB Risk Group were reflexed to Truenat MTB-RIF Dx testing, 38 of which were RIF-resistant as determined by DST. Truenat MTB-RIF Dx had somewhat lower sensitivity than Xpert (difference = -2.6 [95%CI -13.5, +6.8]), again with high uncertainty around this estimate based on one additional false-sensitive test result. As expected, differences in sensitivity were largely driven by differences in sensitivity among the smear-negative culture-positive subgroup. There was no difference in the point estimates for specificity between the two assays.

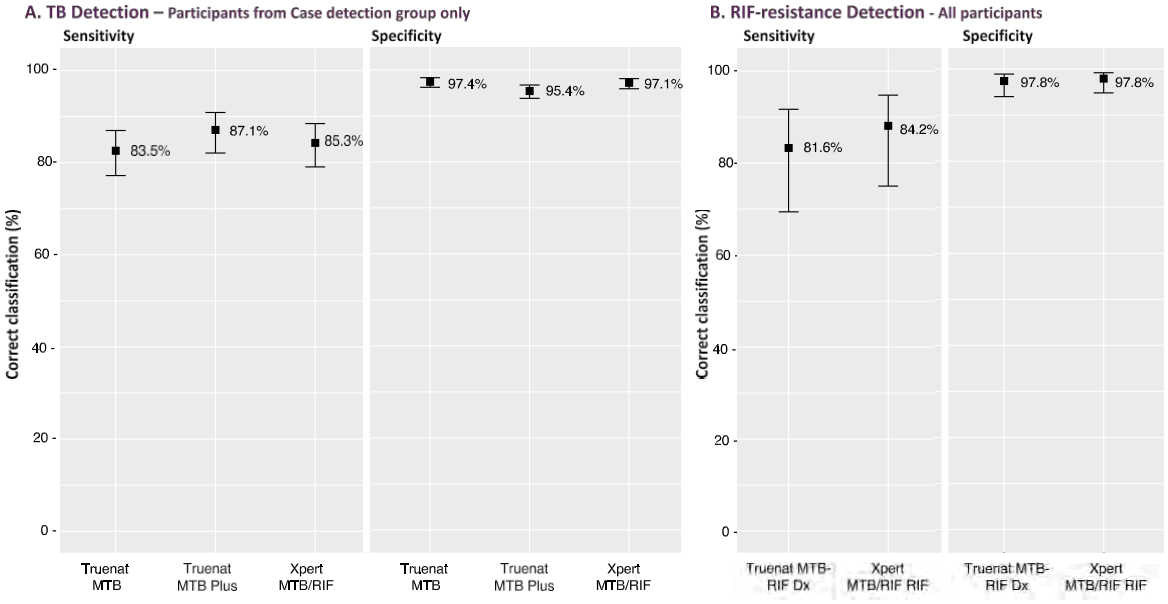


Figure 7: Performance of the Truenat assays and Xpert

Note: A) Diagnostic accuracy for TB detection, compared to Xpert. Only participants in the Case Detection Group are included. B) Performance for RIF-resistance detection compared to Xpert. Participants in both the Case Detection and the

DR-TB Risk Group were included. Analysis was done on all enrolled participants except those enrolled in Peru where Ultra testing was conducted. Squares represent point estimates and bars represent 95% CI.

Table 10: Performance of Truenat assays for TB and RIF-resistance detection compared to Ultra

		N	TP	FP	FN	TN	Sensitivity % (95% CI)		Sensitivity % Smear Pos (95% CI) - N		Sensitivity % Smear Neg (95% CI) - N		Specificity % (95% CI)	
Case Detection Group only	**Truenat MTB**													
	Ultra	377	88	8	4	277	95.7	[89.3,98.3]	100	[93.0,100.0] - N:51	90.2	[77.5,96.1] - N:41	97.2	[94.6,98.6]
	Truenat MTB	377	67	2	25	283	72.8	[63.0,80.9]	94.1	[84.1,98.0] - N:51	46.3	[32.1,61.3] - N:41	99.3	[97.5,99.8]
	Difference (Truenat MTB - Ultra)						-22.9	[-32.4,-15.4]	-5.9	[-15.9,+1.5]	-43.9	[-59.0,-29.9]	2.1	[+0.7,+4.5]
	Truenat MTB Plus													
	Ultra	377	88	8	4	277	95.7	[89.3,98.3]	100	[93.0,100.0] - N:51	90.2	[77.5,96.1] - N:41	97.2	[94.6,98.6]
	Truenat Plus MTB	377	73	7	19	278	79.3	[70.0,86.4]	96.1	[86.8,98.9] - N:51	58.5	[43.4,72.2] - N:41	97.5	[95.0,98.8]
	Difference (Truenat MTB Plus - Ultra)						-16.4	[-25.2,-10.1]	-3.9	[-13.2,+3.4]	-31.7	[-47.0,-19.6]	0.3	[-1.8,+2.6]
Case Detection and DR-Risk Group	**RIF detection**													
	Ultra Rif	57	8	2	0	47	100	[67.6,100.0]	100	[61.0,100.0] - N:6	100	[34.2,100.0] - N:2	95.9	[86.3,98.9]
	Truenat Rif	57	8	2	0	47	100	[67.6,100.0]	100	[61.0,100.0] - N:6	100	[34.2,100.0] - N:2	95.9	[86.3,98.9]
	Difference (Truenat Rif - - Ultra Rif)						0	[-32.4,+32.4	0	[-39.0,+39.0]	0	[-65.8,+65.8]	0	[-7.3,+7.3]

Note: Sensitivity/specificity for Truenat MTB and Truenat MTB Plus for TB detection (reference standard: MTB culture) and sensitivity/specificity for Truenat MTB-RIF Dx for RIF-resistance detection (reference standard: RIF DST); Only participants in the Case Detection Group were include in the TB detection analyses for Truenat MTB and MTB Plus (n=377), whereas for RIF resistance detection analyses participants in both the Case Detection Group and the DR-TB Risk Group were included (n=57). Data are shown for the reference lab sputum only. Truenat performance comparisons were drawn against Ultra for participants from Peru only, the sole study site where Ultra testing was conducted; Among the 377 participants, 41 were smear-negative, culture-positive and 51 were smear-positive, culture-positive.

In Peru, where the reference laboratory used Ultra as the comparator, we observed lower sensitivity for both Truenat MTB (difference = -22.9% [95%CI -32.4, -15.4]) and Truenat MTB Plus (difference = -16.4% [95%CI -25.2, -10.1]) than for Ultra (Table 10 and Figure 8) and this difference was driven by the smear-negative subgroup. Truenat MTB specificity was slightly higher than that of Ultra (difference = +2.1% [95%CI +0.7, +4.5]), whereas the point estimate for Truenat MTB Plus specificity was similar to that of Ultra (difference = +0.3% [95%CI -1.8, +2.6]).

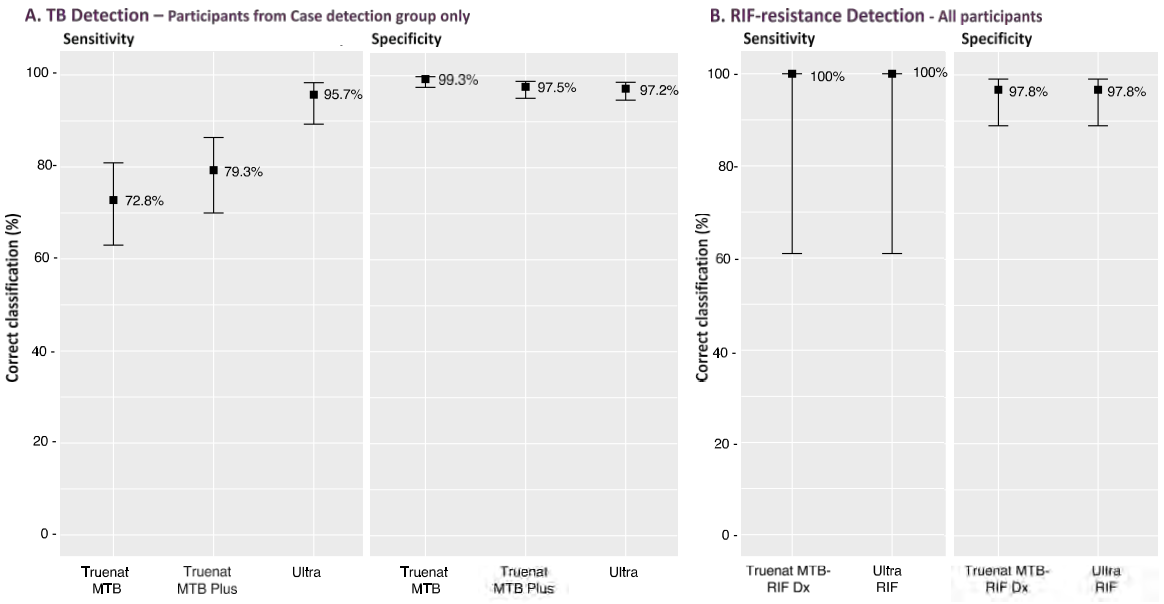


Figure 8: Performance of Truenat assays and Ultra

Note: A) Diagnostic accuracy for TB detection, compared to Ultra. Only participants in the Case-Detection Group are included. B) Diagnostic accuracy for RIF-resistance detection compared to Ultra. Participants in both the Case Detection and the DR-TB Risk Group were included. Analysis was done exclusively on data from participants enrolled in Peru where only Ultra testing was conducted. Squares represent point estimates and bars represent 95% CI.

Truenat MTB Plus performed with increased sensitivity over Truenat MTB. Specificity of Truenat MTB Plus was lower than that of Truenat MTB.

Diagnostic accuracy of the Truenat assays among people living with HIV

In total, only ten TB patients living with HIV were enrolled, therefore formal evaluation of Truenat performance in this subgroup was not done for this analysis.

Effect of TB history on specificity for MTB detection
Varying specificity of molecular assays among people with vs. without a prior history of TB has previously been observed. Table 11 and Table 12 show how the specificity of Truenat assays varies between participants with and without a history of prior TB disease. Xpert is included for comparison purposes in Table 11, and Ultra (participants in Peru only) is shown in Table 12. As seen for Xpert and Ultra, specificity of all Truenat assays was lower in participants with a history of TB disease.

Table 11: Specificity of the Truenat assays compared to Xpert among participants with and without a prior history of TB

	All samples	Specificity % - TB History (95% CI)	Specificity % - No TB History (95% CI)
Case Detection Group only	**Truenat MTB**		
	Xpert MTB/RIF	90.6 [81.0,95.6]	97.6 [96.2,98.5]
	Truenat MTB	92.2 [83.0,96.6]	97.7 [96.4,98.6]
	Difference (Truenat MTB - Xpert)	+1.6 [-5.6,+9.4]	+0.1 [-0.9,+1.2]
	Truenat MTB Plus		
	Xpert MTB/RIF	90.2 [80.2,95.4]	97.5 [96.1,98.4]
	Truenat Plus MTB	88.5 [78.2,94.3]	95.9 [94.2,97.2]
	Difference (Truenat Plus MTB - Xpert)	-1.7 [-11.7,+8.2]	-1.6 [-3.1,-0.2]

Note: Differences in sensitivity and specificity were calculated as performance of each Truenat assay minus Xpert for the reference lab sputum relative to MTB culture (for TB detection). Only participants in the Case Detection Group were included in the TB detection analyses for Truenat MTB and MTB Plus. Truenat performance comparisons were drawn against Xpert for participants from all sites except those enrolled in Peru where Ultra testing was conducted.

Table 12: Specificity of the Truenat assays compared to Xpert Ultra among participants with and without a prior history of TB

	All samples	Specificity % - TB History (95% CI)	Specificity % - No TB History (95% CI)
Case Detection Group only	**Truenat MTB**		
	Ultra	92.9 [85.4,96.7]	99 [96.5,99.7]
	Truenat MTB	97.6 [91.8,99.4]	100 [98.2,100.0]
	Difference (Truenat MTB - Ultra)	+4.7 [+0.2,+11.5]	+1.0 [-0.9,+3.5]
	Truenat MTB Plus		
	Ultra	92.9 [85.4,96.7]	99 [96.5,99.7]
	Truenat Plus MTB	92.9 [85.4,96.7]	99.5 [97.2,99.9]
	Difference (Truenat Plus MTB - Ultra)	0 [-6.2,+6.2]	+0.5 [-1.9,+3.1]

Note: Differences in sensitivity and specificity were calculated as performance of each Truenat assay minus Ultra for the reference lab sputum relative to MTB culture (for TB detection). Only participants in the Case Detection Group were included in the TB detection analyses for Truenat MTB and Truenat MTB Plus. Truenat performance comparisons were drawn against Ultra for participants enrolled in Peru only.

Truenat performance in microscopy centres as compared to centralized reference laboratories
The placement of Truenat equipment in microscopy centres allowed us to evaluate the system in the intended setting of use, whereas testing in reference laboratories allowed us to compare Truenat assay performance directly to that of Xpert and Ultra. We also computed differences in performance for Truenat testing in microscopy centres vs. reference laboratories when testing specimens from the same participant (although not the same specimens were tested).

Table 13: Comparative performance of the Truenat assays performed in the clinics and the reference laboratories

		N	TP	FP	FN	TN	Sensitivity % (95% CI)	Sensitivity % Smear Pos (95% CI) -N	Sensitivity % Smear Neg (95% CI) -N	Specificity % (95% CI)
Case Detection Group only	**Truenat MTB**									
	Ref Lab sputum	1356	200	23	54	1079	78.7 [73.3,83.3]	97.1 [93.4,98.8] -N:172	40.2 [30.3,51.1] - N:82	97.9 [96.9,98.6]
	Microscopy Centre sputum	1356	186	22	68	1080	73.2 [67.5,78.3]	90.7 [85.4,94.2] -N:172	36.6 [27.0,47.4] - N:82	98 [97.0,98.7]
	Difference (Microscopy Centre - Ref lab)						-5.5 [-10.2,-1.2]	-6.4 [-11.5,-2.3]	-3.6 [-14.1,+6.6]	+0.1 [-0.9,+1.1]
	Truenat MTB Plus									
	Ref Lab sputum	1293	212	40	42	999	83.5 [78.4,87.5]	98.3 [95.0,99.4] -N:173	51.9 [41.1,62.4] - N:81	96.2 [94.8,97.2]
	Microscopy Centre sputum	1293	204	37	50	1002	80.3 [75.0,84.7]	96.5 [92.6,98.4] -N:173	45.7 [35.3,56.5] - N:81	96.4 [95.1,97.4]
	Difference (Microscopy Centre - Ref lab)						-3.2 [-7.5,+1.0]	-1.8 [-5.6,+1.6]	-6.2 [-17.7,+5.2]	+0.2 [-1.1,+1.7]
Case Detection and DR-Risk Group	**Truenat RIF**									
	Ref Lab sputum	172	14	6	3	149	82.4 [59.0,93.8]	81.2 [57.0,93.4] - N:16	100 [20.6,100.0] - N:1	96.1 [91.8,98.2]
	Microscopy Centre sputum	172	15	8	2	147	88.2 [65.7,96.7]	87.5 [64.0,96.5] - N:16	100 [20.6,100.0] - N:1	94.8 [90.1,97.4]
	Difference (Microscopy Centre - Ref lab)						+5.8 [-13.6,+27.0]	+6.3 [-14.3,+28.3]	0 [-79.3,+79.3]	-1.3 [-5.0,+1.8]

Note: Differences in sensitivity and specificity were calculated as performance of each Truenat assay conducted in the microscopy centre (Day 2) minus that conducted in the reference lab (Day 1), relative to MTB culture (for TB detection) or RIF DST (for RIF resistance detection); Only participants in the Case Detection Group were included in the TB detection analyses for Truenat MTB and MTB Plus, whereas participants in both the Case Detection Group and the DR-TB Risk Group were included for the analysis of RIF resistance detection. Comparison of Truenat performance in the microscopy centre vs. the reference lab is not a direct head-to-head comparison as different sputa from the same participant were used (Day 1 sputum in the reference lab and Day 2 sputum in the microscopy centre).

For MTB detection, we observed lower sensitivity in the microscopy centre than in the reference laboratory for Truenat MTB (difference = -5.5 [95%CI -10.2, -1.2]), and for Truenat MTB Plus (difference -3.2 [95%CI -7.5, +1.0]) (Table 13). However, these differences could be due to random variability caused by to the known day-to-day fluctuation in bacillary load observed in sputum specimens. There was no appreciable difference in Truenat specificity between sputa run in the microscopy centres and the reference laboratories. Given that the Truenat assays have an open-system format, with the requirement to transfer 6µL of DNA eluate to a qPCR machine, these results are of particular importance. We observed two additional false-resistant results and thus slightly lower specificity for Truenat MTB-RIF Dx in the microscopy centre, although the sample size was limited for RIF resistance in the microscopy centres and thus estimates of the difference uncertain.

In addition to the analyses above, we analysed performance of reference laboratory testing on raw sputum versus on sediment, although testing sediment is not part of the intended use of the Truenat assays. Sensitivity for MTB detection in sediment was lower than that seen for raw sputum for both Truenat MTB (difference = -9.7% [95%CI -13.9, -6.0]), and Truenat MTB Plus (difference = -5.4% [95%CI -9.2, -2.1]). Specificity of MTB detection in sediment was similar to that of sputum using both Truenat MTB (difference = +0.6 [95%CI -0.1, +1.6]), and Truenat MTB Plus (difference = 0% [95%CI -1.4, +1.4]) (Appendix).

Results from QC testing: Daily negative control testing and weekly swab testing

As part of study procedures, sites performed daily negative control and weekly swab testing of both the Truelab and Trueprep machines, as described in the methods (Section 2.1). Overall, positive results from testing swabs and negative controls were rare, indicating appropriate daily cleaning and handling of materials (Table 14). Negative controls and swabs with positive results were less commonly observed in microscopy centres than in reference laboratories. Additionally, all positive results were resolved after cleaning, and did not persist or inhibit subsequent specimen testing. Most importantly, days where swabs or negative controls tested positive never coincided with days where participant specimens tested false-positive, suggesting that the risk of carry-over contamination was low in the context of this study.

Table 14: Proportion of Truenat MTB Plus positive results for daily negative control and weekly swab tests at each site operating the Truelab system.

		Negative controls testing positive		Swabs testing positive	
		(%)	n/N	(%)	n/N
Hinduja	Mumbai Ref Lab (01)	1.2%	2/171	2.9%	1/34
Guwahati	Guwahati Ref Lab	0.0%	0/129	3.0%	4/131
	DTC Kamrup (02)	0.0%	0/106	0.0%	0/18
	Railway (03)	0.0%	0/52	0.0%	0/13
	Sonapur (04)	0.0%	0/40	0.0%	0/7
Chennai	Chennai Ref Lab	0.8%	1/120	0.0%	0/50
	Ayanavaram (05)	0.0%	0/63	0.0%	0/12
	Villiwakkam (06)	0.0%	0/26	0.0%	0/4
	Thanthai Perivar (07)	0.0%	0/88	0.0%	0/18
Ahmedabad	Ahmedabad Ref Lab	1.3%	4/293	3.2%	5/151
	Madhupura (08)	0.0%	0/117	0.0%	0/24
	CHC Chhala (09)	3.3%	3/87	0.0%	0/20
	PHC Kuha (10)	1.9%	2/104	4.0%	1/24
Peru	Lima - UPCH Ref Lab	0.0%	0/100	0.0%	0/107
	CS Huascar II (11)	0.0%	0/35	0.0%	0/35
	CS Huascar XV (12)	0.0%	0/33	0.0%	0/39
	CS Jose Carlos Mariategui (13)	0.0%	0/32	0.0%	0/43
	CS Fraternidad (14)	0.0%	0/51	0.0%	0/96
	CS El Porvenir (19)	0.0%	0/50	0.0%	0/40
Ethiopia	Addis Ababa - EPHI Ref Lab	1.8%	3/162	3.8%	2/50
	Hiwot Amba (15)	0.0%	0/21	0.0%	0/12
	St. Gebrel (16)	0.0%	0/20	0.0%	0/5
	Woreda 01 (17)	0.0%	0/34	0.0%	0/10
PNG	Port Moresby: CPHL (18)	3.3%	2/59	4.8%	1/20

Note: Truenat MTB Plus was used for all negative control and swab testing procedures. Negative controls were run daily at each site, and swabs were tested weekly at each site. Data presented exclude all non-determinate results.

Non-determinate results on the Truenat system

Non-determinate results were excluded from analysis of sensitivity and specificity and are reported separately in this section. Table 15 provides and overview of the proportion of non-determinate results for the Trueprep extraction, the Truenat assays as well as Xpert and Ultra assays as comparators.

Table 15: Proportion of non-determinate results for Trueprep extraction, Truenat assays and Xpert and Ultra assays

Total non-determinates	Initial Test		Repeat Test	
	(%)	n/N	(%)	n/N
Trueprep	2.4%	113/4732	11.7%	13/111
Truenat MTB	6.2%	293/4720	21.2%	62/293
Truenat MTB Plus	9.2%	434/4720	36.8%	159/432
Truenat MTB RIF-Dx*	22.5%	232/1042	72.7%	157/216
Xpert MTB/RIF	2.6%	65/2522	7.9%	5/63
Xpert Ultra	0.0%	0/786	-	-

Note: Non-determinate results represent a combination of operator and equipment errors or failures, invalid results and indeterminate results, for all participant specimens tested as part of this study. The non-determinate results for the Truelab micro PCR machine represent results for all different Truenat assay performed at each site. Not all specimens that failed on the initial test were still available for repeat testing. The results presented here do not capture errors in DNA loading or chip loading as site incident logs did not report high levels of such errors. The full table showing the proportion of individual assay non-determinate test from each site is presented in the Appendix. The proportion of study participants affected by non-determinate initial and repeat tests is shown in Figure 10 below. * Note that the Truenat MTB-RIF Dx was run on any specimen that tested positive for MTB by either the Truenat MTB assay or the Truenat MTB Plus assay. See more detailed results below in Table 16.

Trueprep non-determinate results

The proportion of initial Trueprep non-determinate results was 2.4% (Table 15). Repeat testing resolved results for 88.3% specimens that failed on the initial test. The manufacturer has indicated that sputum may be stored in lysis buffer for up to 48 hours with no degradation of DNA, allowing time for repeat testing if enough specimen is available.

Two reference labs were disproportionately affected by a high rates of Trueprep failures: The reference lab in Ethiopia had 5.4% non-determinate Trueprep results (21 of 387 tests), and the reference lab in Guwahati, India had 6.4% non-determinate Trueprep results (31 of 487 test) (Appendix). Repeat testing resolved 90% of participant samples in Ethiopia and 78% in Guwahati. The root cause behind these Trueprep errors was related to operators not exchanging DNA extraction kit buffers appropriately.

Non-determinate results for the Truenat MTB and Truenat MTB Plus assays

An analysis of initial and repeat Truenat invalid results is presented in Table 15 and Appendix. Initial test non-determinate proportions for the Truenat MTB and MTB Plus chip were 6.2% and 9.2%, respectively. Of the tests that failed, 21.2% and 36.8% remained non-determinate upon repeat testing. Comparatively, the non-determinate rate of Xpert was 2.6%, with no failures observed for Ultra. A detailed analysis of the frequency of different error types is provided in the Appendix.

Non-determinate results for the Truenat MTB-RIF Dx assay

In comparison, the non-determinate rate for the Truenat MTB-RIF Dx assay initial test was 22.5%, of which 72.7% did not resolve upon repeat testing. The non-determinate rate varied strongly depending on the bacterial load in the specimen: the proportion of non-determinate Truenat MTB-RIF Dx results was 6.7% if reflexed from a Truenat MTB-positive result vs. 72.2% if reflexed from a Truenat MTB-negative / Truenat MTB Plus-positive specimen (
Table 16).

Table 16: The proportion of non-determinate Truenat MTB-RIF Dx results when reflexed from either the Truenat MTB or MTB Plus TB detection result

	Truenat RIF-Dx Non-determinates	
	% (95% CI)	n/N
If reflexed from Truenat MTB-pos and MTB Plus-pos	3.9% (2.7, 5.4)	32/830
If reflexed from Truenat MTB-neg and MTB Plus-pos	67% (60, 74)	120/179
If reflexed from Truenat MTB-pos and MTB Plus-neg	72% (56, 84)	26/36
If reflexed only from Truenat MTB-pos	6.7% (5.2, 8.6)	58/866
If reflexed only from Truenat MTB-Plus-pos	15% (13, 17)	152/1009

Greater resolution of the specific types of errors seen on the Truelab and Trueprep devices is provided in the Appendix.

Non-determinate results by site, specimen and assay type

We also evaluated the proportion of non-determinate Truenat results by specimen type, site and location. Results for the Truenat MTB Plus are shown in Figure 9 (figures for other chips are included in the Appendix). The non-determinate rates in Ethiopia were higher than the median across all sites (dotted red line). A high number of indeterminates clustered over a 2-week period in July 2019, on two Truelab machines in the reference lab. These machines were replaced with backup machines. The proportion of non-determinates decreased in August and September 2019, suggesting the initial occurrence may have been equipment-related or operator-related. A site technical visit conducted in August failed to identify the root cause, and the equipment appeared functional. The most likely cause

of these errors was determined to be wash buffer unavailability on the Trueprep machine (empty bottle, tube blockage, incorrectly sealed bottle or failure to pressurize).

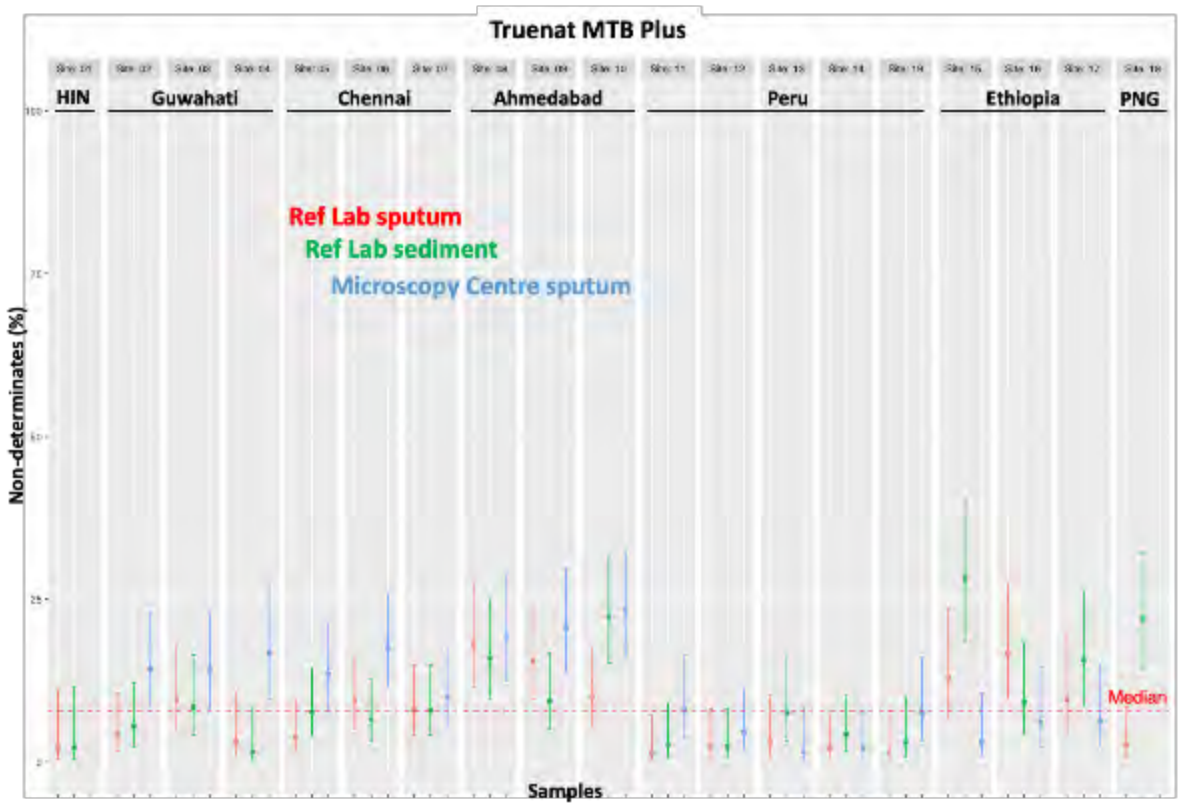


Figure 9: Proportion of Truenat MTB Plus chip indeterminate results at each site, for each specimen type
Non-determinates results included invalid, indeterminate and error results on the Truelab PCR machine. The dashed red line represents the median proportion of indeterminate results for all sites. Sputum specimen types are presented as coloured bars, with dots representing point estimates and bars representing 95% CI. Each site is indicated by grey shading. Figures for the Truenat MTB and the Truenat MTB-RIF Dx chips can be found in the Appendix.

Resolution of non-determinate Truenat results upon repeat testing

Upon re-testing of specimens in the microscopy centre, more than two-thirds of all Truenat MTB or MTB Plus non-determinate results resolved, leading to a non-determinate rates of 1.7% and 3.9% for MTB and MTB Plus, respectively, when allowing for a single repeat test (Figure 10).
In contrast, only 28% of the RIF resistance detection chips that were non-determinate on initial testing resolved. This was likely due to the lower sensitivity issues of Truenat MTB assay than Truenat MTB Plus, as discussed above.

Comparatively, the rate of initial Xpert non-determinates was 2.6% (65/2522 tests) with no non-determinate results for Ultra; 92% Xpert non-determinates resolved upon repeat testing.

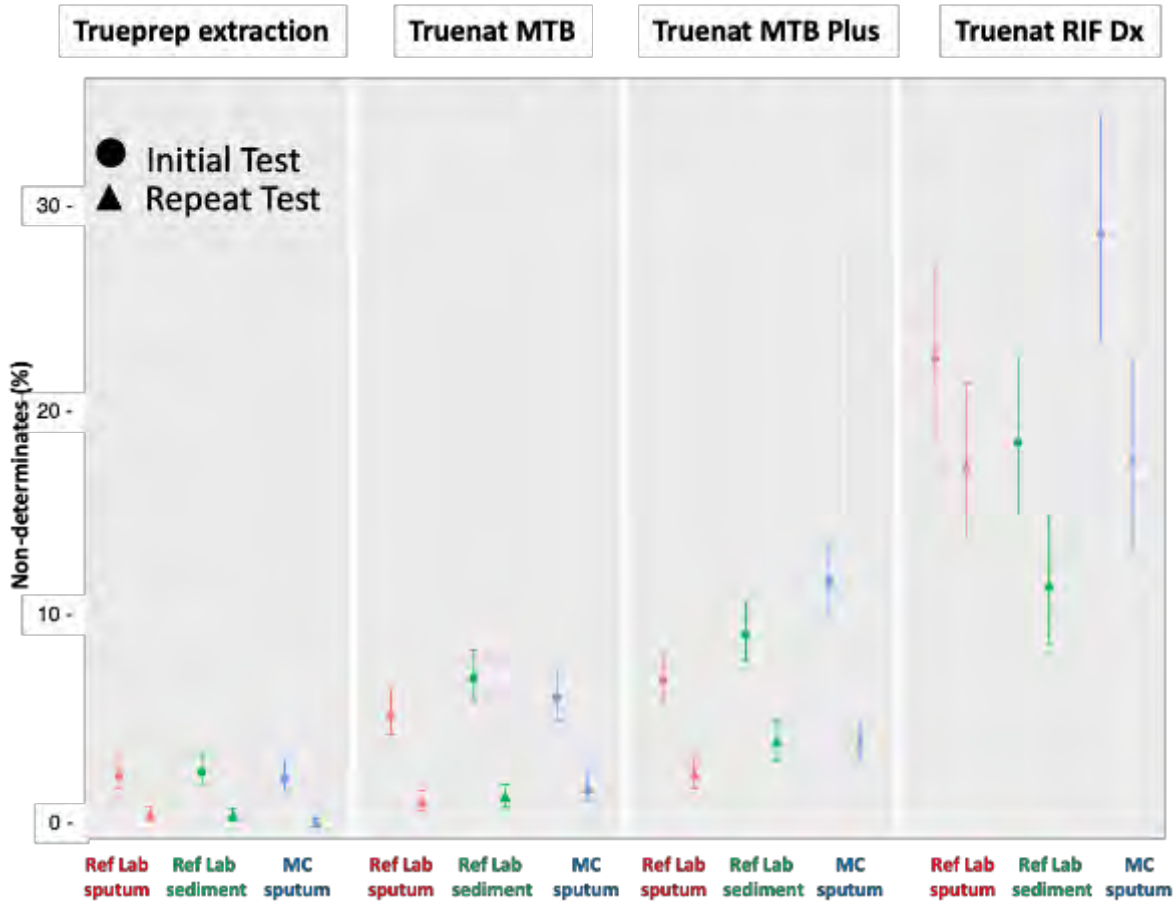


Figure 10: The proportion of participants with non-determinate Trueprep or Truenat assay results upon initial and repeat tests:

The proportion of participants with non-determinate Trueprep or Truenat assay results upon initial and repeat tests: Non-determinate results included invalid, indeterminate and error results on the Trueprep and Truelab PCR machine. Sputum specimen types are presented as coloured bars. The centre dots or triangles represents the point estimate for initial and repeat tests, and bars represent 95% CI. Each Truenat MTB assay type is indicated by grey shading. MC = Microscopy Centre.

Root cause analysis for non-determinate results

Root cause analysis revealed that the vast majority of Truenat assay invalid results can be traced back to a Trueprep operational issue during the DNA extraction processing steps. Upon notification by sites of a series of consecutive invalid Truenat results, root cause analysis and corrective and preventive actions were implemented. We identified the most likely cause of Truelab non-determinates PCR results was the unavailability of Trueprep buffers during the DNA extraction process, either due to failure of the operator to replace Trueprep buffers, incorrect attachment of buffer bottles to equipment, or tube blockage. Under regular operation, the Trueprep machine signals to the operator to replace buffers after every set of 25 extractions. In some cases, the operator had disabled the warning alarm and failed to replace buffer bottles for the Trueprep machine. In other circumstances, some buffers may not have had sufficient volume to complete 25 runs or the bottle was not correctly loaded after replacement, and therefore led to errors. As the DNA eluate buffer was provided in excess (per test) to that of the wash buffers, in many such scenarios the DNA eluate was still available at the end of the run but may have contained PCR inhibitors because of incomplete washing steps. This subsequently led to consecutive Truenat PCR invalid results.

Discordance analysis

A list of all specimens with discordant results between Truenat assays and culture is provided in the appendix.

MTB Detection Assays

Overall there were 126 participants who had at least one false-positive result on at least one of the six tests done per participant (each participant had 3 specimens tested with two assays). Similarly, 131 participants with at least one false-negative result by either Truenat MTB or MTB Plus assays were identified.
Of the 126 participants with false-positive results, 40 were also false-positive by either Xpert or Ultra. Of the remaining 86 false-positive results, 12 had a prior history of TB and for 11 their TB history was unknown.
None of the false-positive results coincided with positive test results on negative controls or swabs. Most false-positive results were observed from tests done in the reference laboratories. Specifically, the proportion of participants with a false-positive Truenat result when tested in the microscopy centre from either initial or repeat tests was 7.9% (9/114) for Truenat MTB and 17.5% (20/114) for Truenat MTB Plus, the remainder arose from testing in the reference laboratory.

Of the 131 participants with a false-negative Truenat MTB or MTB Plus result on any specimen tested, 55 participants were negative by Truenat MTB and 37 were negative by Truenat MTB Plus on all sputum specimens tested with valid results; 34 were also missed by Xpert and/or Ultra on both direct sputum and decontaminated sediment. 22 participants were missed by all 3 molecular assays, for which one participant was smear-positive and three had a history of TB.
The remaining 93 were positive by at least one Truenat test on at least one sputum specimen.
93% (n = 122) of participants with any false-positive Truenat result were smear-negative or scanty. Of the 39 participants with smear-positive or scanty results, 87% were detected by Truenat MTB, 95% were detected by Truenat MTB Plus and 95% were detected Xpert or Ultra on at least one specimen.

RIF Resistance assay
There were 10 participants with RIF-resistant results based on Truenat MTB-RIF Dx that tested RIF-sensitive on phenotypic DST. Of these, six participants also tested RIF-resistant on Xpert or Ultra. Five of these six Truenat MTB-RIF Dx-resistant participants tested as RIF-resistant on both of the specimens provided and tested on two separate days in two separate locations.

There were 10 participants with RIF-sensitive results based on Truenat MTB-RIF Dx that tested RIF-resistant on phenotypic DST. Of these, six participants also tested RIF-sensitive on Xpert or Ultra. Two of these six Truenat MTB-RIF Dx-sensitive participants tested as RIF-sensitive on both of the specimens provided and tested on two separate days in two separate locations. The other participant had only a single valid RIF result (sensitive).

Additional testing
Additional laboratory testing for cases with discordant results (and an equal number of non-discordant cases) is currently underway, as described in methods section 2.1 and according to an SOP developed prior to study start. In brief, we will use targeted sequencing to identify the presence of MTB and NTMs in stored DNA eluate and in remaining sputum or sediment from participant specimens, where available. Where appropriate, we will also identify off-target amplification in amplicons taken directly from positive Truenat chips stored at each site.

Discussion

We performed a multicentre diagnostic test accuracy study to evaluate the performance of the point of care Truenat MTB Detection assays and RIF-resistance assay in the intended setting of use and in this report provide analyses based on completed enrolment and all available culture results at the time of writing. The final analysis will add culture results from 98 participants from PNG. Over-enrolment from the trial in total has ensured that analyses presented here are adequately powered for the primary objectives.

Overall, the findings suggest that the Truenat assays for MTB detection have good performance characteristics and could be considered as initial test for the diagnosis of TB. The primary analyses focused on performance in the microscopy centre setting and suggested good performance of the assays. The sensitivity of Truenat MTB and MTB Plus in the microscopy centre was estimated to be 73% (95%CI 68, 78) and 80% (95%CI 75, 84), respectively. Specificity of the Truenat assays was 98% (95%CI 97, 99) and 97% (95%CI 95, 97). We observed that specificity of the assays in the microscopy centre was equivalent to that seen in the reference lab, with only marginal difference in sensitivity for each assay between tests done in the microscopy centres and the reference laboratories. Comparative data (testing the same specimens side by side) on Truenat and Xpert assays was available from testing in the reference lab and suggested overall similar performance. The sensitivity of Xpert was higher than of Truenat MTB but lower than that of Truenat MTB Plus. Sensitivity of Ultra was higher than that of both Truenat MTB and MTB Plus. The specificity of Truenat MTB was similar to that of Xpert, and the specificity of Truenat MTB Plus was similar to that of Ultra.
The proportion of non-determinate Trueprep results on initial testing was 2.4% (113/4732 test), with almost half of these arising from one operator on one machine at one reference laboratory. Upon retesting 87.5% of non-determinate results resolved. The proportion of non-determinate Truenat MTB and MTB Plus assay results on the initial test was 6.2% and 9.2%, respectively; this resulted in 6.2% of participants not having a valid Truenat MTB results and and 11.8% no valid MTB Plus assay result on specimens tested in the microscopy centre on initial test. When allowing for a single repeat-test in the microscopy centre, only 1.7% and 3.9% of participants remained with non-determinate results for Truenat MTB and MTB Plus, respectively. This is largely in line with that for Xpert non-determinate results (2.6%) although unlike Xpert, the Molbio assays have been conducted in primary health care facilities.

Additional analyses performed in subgroups and specimens tested outside the microscopy centre setting support the overall good performance of the Truenat MTB and MTB Plus assays. Sensitivity of the Truenat MTB and MTB Plus assays on smear-positive culture positive specimens at the microscopy centre was 91% and 96%, and amongst smear-negative culture-positive participants sensitivity was 37% and 46%, respectively. Specificity of the Truenat assays in the microscopy centre setting and the reference laboratories was equivalent. As previously observed for other molecular assays, Truenat assay specificity was reduced in individuals with a prior history of TB disease, and this was more pronounced for the Truenat MTB Plus and Ultra assays than for Truenat MTB and Xpert. These results are to be expected given the higher sensitivity of these assays compared to their counterparts, as each assay may also detect minimal amounts of non-viable or non-culturable bacilli. Sensitivity for MTB detection in sediment was lower than that seen for raw sputum for both Truenat MTB and Truenat MTB Plus, while specificity was similar. However, as the assays are intended for use on unprocessed sputum, not sediment (decontaminated sputum), this is not expected to be relevant in practice and indeed the manufacturer does not list sediment as a specimen type in the instructions for use.

Data available on the performance of the Truenat MTB-RIF Dx assay was limited. The Truenat MTB-RIF Dx assay for detection of RIF resistance had a sensitivity of 84% (95%CI 62, 95) and specificity of 95% (95%CI 91, 98) in the microscopy centre. However, most RIF-resistant participants were enrolled at PD Hinduja Hospital in Mumbai, India, a DR-TB referral centre that does not have a separate microscopy centre. Thus, these results have high uncertainty due to the small sample size (only 19 RIF-resistant participants were enrolled and tested by Truenat MTB-RIF Dx in the

microscopy centre setting). Evaluation of RIF-resistance detection on sputum tested in the reference lab (where specimens from 51 RIF-resistant and 258 RIF-sensitive participants provided valid results) showed sensitivity of 84% (95%CI 72, 92]) and specificity of 97% (95%CI 95, 99). Sensitivity and specificity of both Xpert and Ultra for detection of RIF-resistance was similar to that of Truenat MTB-RIF Dx when tested on the same specimens. To complement the clinical data that will become available by completion of the study, FIND and NIRT have completed an analytical study testing a strain panel containing 90% of all global RIF-resistance mutations. The Truenat MTB RIF-Dx assay detects >90% of the global prevalence of RIF-R mutations.

The proportion of non-determinate results for the Truenat MTB-RIF Dx assay was high. Of 1,045 Truenat MTB RIF-Dx initial tests run, 20% of all initial tests run were non-determinate, and 73% of these remained unresolved upon re-testing. We found that the Truenat MTB-RIF Dx assay non-determinate rate varied heavily depending on the bacillary load in the specimen. Overall, the proportion of non-determinate Truenat MTB-RIF Dx results was 6.7% if reflexed from a Truenat MTB-positive result. In contrast, the non-determinate rate was 72% if reflexed from a specimen that tested positive only on 'Truenat MTB Plus-positive' (i.e. was 'Truenat MTB-negative'). This indicates that the increased sensitivity of Truenat MTB Plus to detect MTB is likely higher than that of the Truenat MTB-RIF Dx chip to detect RIF resistance, thereby producing a high number of indeterminate RIF resistance results. This is similar to scenarios in which Ultra trace results do not provide a corresponding RIF resistance result. Semi-quantitative results (e.g. a Ct value cut-off) for Truenat MTB Plus assay could be considered when deciding whether performing a Truenat MTB-RIF Dx reflex test would be worthwhile or likely to yield non-determinate results. Further root cause analysis revealed some key underlying reasons for non-determinate results that may be prevented in the future: errors made during use of the Trueprep DNA extraction led insufficient washing and subsequently to increased error rates during PCR.

This study and report has some limitations. Firstly, some of our analyses were limited by two factors: Ultra (but not Xpert) was performed in one country, whereas Xpert (but not Ultra) was performed at all other countries and thus we are not able to provide data on direct head-to-head to these assays on all specimens; two sites (PD Hinduja Hospital in Mumbai, India, and the Port Moresby General Hospital in Papua New Guinea) were not able to do testing at the microscopy setting and thus the numbers for analyses at microscopy centre setting were further reduced. Secondly, only ten people living with HIV were included in the analysis to date. Thirdly, while we found specificity to be adequate, stringent cleaning procedures (and daily negative control and weekly swab testing) were implemented as part of the Truenat study described here, and it is possible that performance could differ under more routine conditions if such procedures are not followed.

In conclusion, for MTB detection, the analysis of the prospective clinical study suggests that the Truenat MTB and MTB Plus assays may have similar accuracy to that of Xpert and Ultra and can be performed at microscopy centre level. Non-determinate rates were higher than for Xpert and Ultra. The data for Truenat MTB-RIF Dx detection of RIF-resistance, especially that in the microscopy centre, were limited. The Molbio platforms and assays have the potential to meet the minimal criteria set by the WHO TPP for a smear-replacement test.

References

1. Albert H, Nathavitharana RR, Isaacs C, Pai M, Denkinger CM, Boehme CC. Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? Eur Respir J Off J Eur Soc Clin Respir Physiol. 2016 Aug;48(2):516–25.
2. MacPherson P, Houben RM, Glynn JR, Corbett EL, Kranzer K. Pre-treatment loss to follow-up in tuberculosis patients in low- and lower-middle-income countries and high-burden countries: a systematic review and meta-analysis. Bull World Health Organ. 2014 Feb 1;92(2):126–38.
3. World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting - 2014 [Internet]. 2019 Apr [cited 2019 Nov 5]. Available from: https://apps.who.int/iris/bitstream/handle/10665/135617/WHO_HTM_TB_2014.18_eng.pdf;jsessionid=050A476ED684FF1B92A66BC863C6BA4C?sequence=1
4. World Health Organization. Global Tuberculosis Report 2019.
5. Dye C, Watt CJ, Bleed DM, Hosseini SM, Raviglione MC. Evolution of Tuberculosis Control and Prospects for Reducing Tuberculosis Incidence, Prevalence, and Deaths Globally. JAMA. 2005 Jun 8;293(22):2767–75.
6. World Health Organization. Global Tuberculosis Report 2018. 2018.
7. Molbio Diagnostics Pvt. Ltd. [Internet]. [cited 2019 Oct 25]. Available from: http://www.molbiodiagnostics.com/
8. Nikam C, Kazi M, Nair C, Jaggannath M, M M, R V, et al. Evaluation of the Indian TrueNAT micro RT-PCR device with GeneXpert for case detection of pulmonary tuberculosis. Int J Mycobacteriology. 2014 Sep;3(3):205–10.
9. Nikam C, Jagannath M, Narayanan MM, Ramanabhiraman V, Kazi M, Shetty A, et al. Rapid diagnosis of Mycobacterium tuberculosis with Truenat MTB: a near-care approach. PloS One. 2013;8(1):e51121.
10. Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, et al. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. N Engl J Med. 2010 Sep 1;363(11):1005–15.
11. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database of Systematic Reviews. 2014.
12. World Health Organization. Rapid implementation of the Xpert MTB/RIF diagnostic test: technical and operational “How-to”; practical considerations. 2011.
13. World Health Organization. Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children Policy update. 2013.
14. Helb D, Jones M, Story E, Boehme C, Wallace E, Ho K, et al. Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. J Clin Microbiol. 2010;
15. Chakravorty S, Simmons AM, Rowneki M, Parmar H, Cao Y, Ryan J, et al. The New Xpert MTB/RIF Ultra: Improving Detection of Mycobacterium tuberculosis and Resistance to Rifampin in an Assay Suitable for Point-of-Care Testing . mBio. 2017;
16. Theron G, Peter J, Calligaro G, Meldau R, Hanrahan C, Khalfey H, et al. Determinants of PCR performance (Xpert MTB/RIF), including bacterial load and inhibition, for TB diagnosis using specimens from different body compartments. Sci Rep. 2014;
17. Zeka AN, Tasbakan S, Cavusoglu C. Evaluation of the GeneXpert MTB/RIF assay for rapid diagnosis of tuberculosis and detection of rifampin resistance in pulmonary and extrapulmonary specimens. J Clin Microbiol. 2011;
18. Nicol MP, Workman L, Isaacs W, Munro J, Black F, Eley B, et al. Accuracy of the Xpert MTB/RIF test for the diagnosis of pulmonary tuberculosis in children admitted to hospital in Cape Town, South Africa: A descriptive study. Lancet Infect Dis. 2011;
19. Rufai SB, Kumar P, Singh A, Prajapati S, Balooni V, Singh S. Comparison of Xpert MTB/RIF with line probe assay for detection of rifampin-monoresistant Mycobacterium tuberculosis. J Clin Microbiol. 2014 Jun;52(6):1846–52.

20. Zetola NM, Shin SS, Tumedi KA, Moeti K, Ncube R, Nicol M, et al. Mixed Mycobacterium tuberculosis complex infections and false-negative results for rifampin resistance by GeneXpert MTB/RIF are associated with poor clinical outcomes. J Clin Microbiol. 2014 Jul;52(7):2422–9.
21. Williamson DA, Basu I, Bower J, Freeman JT, Henderson G, Roberts SA. An evaluation of the Xpert MTB/RIF assay and detection of false-positive rifampicin resistance in Mycobacterium tuberculosis. Diagn Microbiol Infect Dis. 2012;
22. Dorman SE, Armstrong DT, King B, Schumacher SG, Nabeta P, Boehme CC, et al. Xpert MTB/RIF Ultra for detection of Mycobacterium tuberculosis and rifampicin resistance: a prospective multicentre diagnostic accuracy study. The Lancet Infectious Diseases. 2018.
23. Gidado M, Nwokoye N, Nwadike P, Ajiboye P, Eneogu R, Useni S, et al. Unsuccessful Xpert® MTB/RIF results: the Nigerian experience. Public Health Action. 2018 Mar 21;8(1):2–6.
24. Lumb R, Van Deun A, Bastian I, Fitz-Gerald M, Pathology SA, Stop TB Partnership (World Health Organization). Laboratory diagnosis of tuberculosis by sputum microscopy: the handbook. 2013.
25. Dharan NJ, Amisano D, Mboowa G, Ssengooba W, Blakemore R, Kubiak RW, et al. Improving the Sensitivity of the Xpert MTB/RIF Assay on Sputum Pellets by Decreasing the Amount of Added Sample Reagent: a Laboratory and Clinical Evaluation. J Clin Microbiol. 2015 Apr 1;53(4):1258–63.
26. UNAIDS Data 2019 [Internet]. [cited 2019 Oct 25]. Available from: https://www.unaids.org/sites/default/files/media_asset/2019-UNAIDS-data_en.pdf

SECTION 3: Assessment of operational characteristics

The aim of the study was to assess the operational characteristics and user appraisal of the MTB Plus and MTB-RIF Dx assays for the detection of pulmonary TB and resistance to RIF under routine conditions at the intended settings of use in India.

Study design

This was a multi-centre operational study carried out at nine microscopy centres and one private laboratory in various geographic locations in India (see Table 17). Most of these laboratories were also subsequently involved in the global evaluation trial (Section 2).

Table 17: List of participating sites

Site	Location	City
Madhupura Urban Health Centre	Urban	Ahmedabad
CHC Chhala	Rural	Ahmedabad
PHC Kuha	Rural	Ahmedabad
DTC Kamrup Metro	Hilly/peri-urban	Guwahati
Sonapur District Hospital	Tribal	Guwahati
Railway Hospital	Hilly/peri-urban	Guwahati
Ayanavaram UPHC	Peri-urban	Chennai
Villiwakkam UPHC	Urban	Chennai
Thanthai Periyar	Peri-urban	Chennai
Dr B Lal Clinical & Molecular Diagnostic Laboratory	Urban	Jaipur

Study training

On-site training was provided by Molbio over two days, in line with how training is provided to customers routinely. Training included: devices setup, hands-on practice by two operators using leftover specimens, Trueprep and Truelab analyzer operation and troubleshooting. Training aids and posters were supplied. FIND provided training on data collection and testing of control swabs to detect the potential for DNA/amplicon contamination within the work environment. Operators were experienced TB laboratory technicians with no experience on molecular tests, except for those at the private laboratory site.

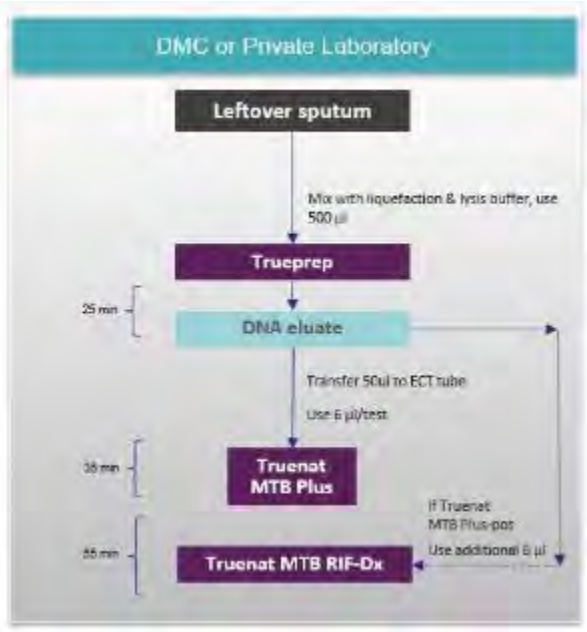


Figure 11: Specimen workflow

Note: No differentiation was done among leftover sputum specimens considered for testing by Truenat i.e. these included either diagnostic or follow-up specimens submitted to the participating labs as per routine

Data collection

Data collection forms and questionnaires were prepared by FIND who trained all participating sites on how to complete these (

Table 18). Data on operational characteristics and user appraisal of the Truenat MTB Plus and RIF-Dx assays, as well as the Trueprep and Truelab devices, was captured throughout the study, i.e. upon setup, during study conduct and at study end.

Table 18: Description of questionnaires used to assess operational characteristics and user appraisal

Questionnaire	Time of completion	Topic(s)
Supervisor (Admin profile user)	End of Training Study end	Ease of use, training, troubleshooting
User set 1 (lab technician)	During training	Setup, user-friendliness, training
User set 2 (lab technician)	Daily	Trueprep & Truelab features, issues
User set 3 (lab technician)	Study end	Overall appraisal

At the end of the study, participating lab technicians were asked to appraise the training process, rate the ease of use, indicate whether they were able to perform certain steps and finally provide an overall appraisal after use over 1 month.

Testing results, including the results of external controls and swabs, were captured in a separate form. Electronic data capture by double data entry was done at NIRT using a dedicated database (OpenClinica).

Analysis

User demographics and user appraisal data was summarized using descriptive statistics. Test results were analyzed both overall and by participating study site where applicable

Results

On-site training took place between July and August 2017 at all sites. Laboratory technicians were trained and processed overall >500 specimens within 4-5 weeks.

User appraisal
A total of 10 laboratory technicians (average 9 years of experience; range 1.5-24 years) participated in the study (Appendix). Moreover, 10 laboratory supervisors (average 9 years of experience; range 2.5-20 years) also provided feedback after training and at study end.

Table 19: Lab technician's appraisal immediately after training (Questionnaire: User set 1):

Aspect	Appraisal
Average setup time (time from when each device was brought out of the case until turned on, in minutes)*	Trueprep: 27 minutes (range 10-40 minutes) Truelab: 26 minutes (range 10-40 minutes)
User friendliness	Trueprep and Truenat considered user friendly 10/10 considered the display to be clear and were able to: i) navigate all screens within applications, ii) read all text and understand terminology used, iii) understand intent of icons on display for both devices
Average number of specimens run before feeling comfortable	4 specimens (range 1-6 specimens)
Training	9/10 considered training to be sufficient, 1/10 considered training to be insufficient 10/10 considered training material to be sufficient (suggestion was made on training videos covering interpretation of results, errors)

*Installation and setup was performed by Molbio as per their current plan for implementation, therefore self-installation/setup was not assessed

Using a daily appraisal questionnaire (Appendix Tables), lab technicians reported both Trueprep and Truelab devices to be easy to start-up. Most sites charged the battery of both devices daily; if the battery was running low, a warning message would appear either before or during a run. No reports were received of any runs being interrupted due to power issues. In 10% (28/280) of cases, the lab technicians reported that the Trueprep device was hot to the touch at a given point during the working day – most of these cases (25/28) were reported in a single lab.

Other features, such as test progress for both devices and the manual steps being prompted on the Trueprep, were generally visible/worked well. Lab technicians reported making one or more technical and operational mistakes on a given day (such as chip loading and difficulty in data entry), 3.5% (10/280) and 22.9% (64/280) when using Trueprep and Truelab, respectively.

Based on the study end appraisal questionnaire, all respondents (10/10) expressed that the devices were user-friendly. Also, all considered that the instructions for cleaning were sufficient and the procedure itself was easy for both Trueprep and Truelab.

Feedback about whether the devices were able to withstand everyday use/handling was overall positive for both Trueprep and Truelab in terms of robustness, battery life, build quality, form factor, size of the device and portability, stability and endurance to environmental conditions (see Appendix).

When asked about which method was considered the most practical or appropriate to transfer the Truelab results to a laboratory information management system (e.g. Nikshay et al), 60% (6/10) indicated that automatic transfer after each run was the preferred method, 30% (3/10) preferred manual transfer once a day and 10% (1/10) preferred manual transfer after each run.

Lab technicians were also asked to provide their overall appraisal at the end of the study. 70% (7/10) of them recommended the use of Truenat. Those who did not recommend it gave the following reasons: i) *pipetting of 6 µL very hard and needs to be perfect*, ii) *additional staff would be required in high workload labs*, iii) *too many invalid results*. When asked about the main barriers for adoption in routine settings (assuming the cost is affordable) lab technicians considered the following characteristics too be important considerations: total duration of the assay run (9/14); manual steps/complexity of the assay (3/14); precision pipetting step (1/14); and low throughput (single test at a time) (1/14).

Input on the main benefits and disadvantages of the assay, as well as the areas where Truenat could be improved, was also provided. The results, including statements (captured by free text), are shown below.

Technical issues
Of a total of 23 devices installed by Molbio (Trueprep and Truelab), 3 (13%) Trueprep devices had to be replaced due to technical issues observed upon installation. Molbio explained that this was likely due to damage during transit. Of these, 2 occurred at the same site consecutively and the devices were replaced over the course of 1 week (Site 1).

Other technical problems reported included: Trueprep cartridge holder failure 3.2% (9/280), Truelab analyzer unable to read the chip's memory 2.1% (6/280) and Truelab chip loading dock failure 7.5% (21/280). In addition, lab technicians reported having consulted the troubleshooting steps and contacted Tech Support for Trueprep in 14% (39/280) and 14.6% (41/280) of cases, respectively, and for Truelab in 16.1% (45/280) and 18.2% (51/280) of cases, respectively.

Truenat testing
A total of 443 sputum specimens were processed by Truenat MTB Plus from 18 July to 11 September 2017. Of these, 384 had valid results for MTB. The overall MTB positivity was 28.1% (108/384). Among 108 MTB positive specimens, 71 had valid results for RIF detection, of which 22.5% (16/71) were RIF-resistant and 77.5% (55/71) were RIF-sensitive.

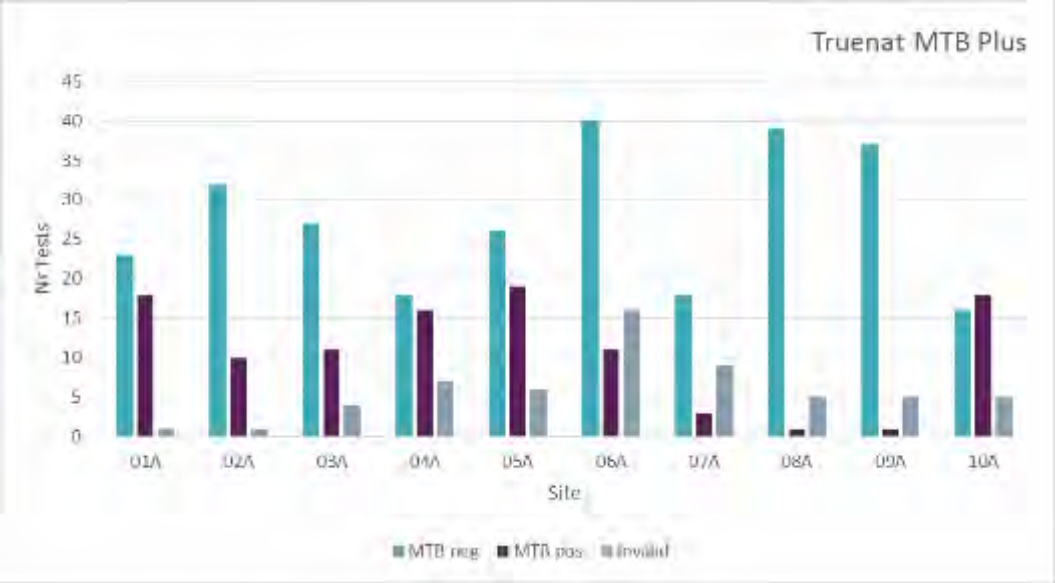


Figure 12: Truenat MTB Plus testing distribution by site

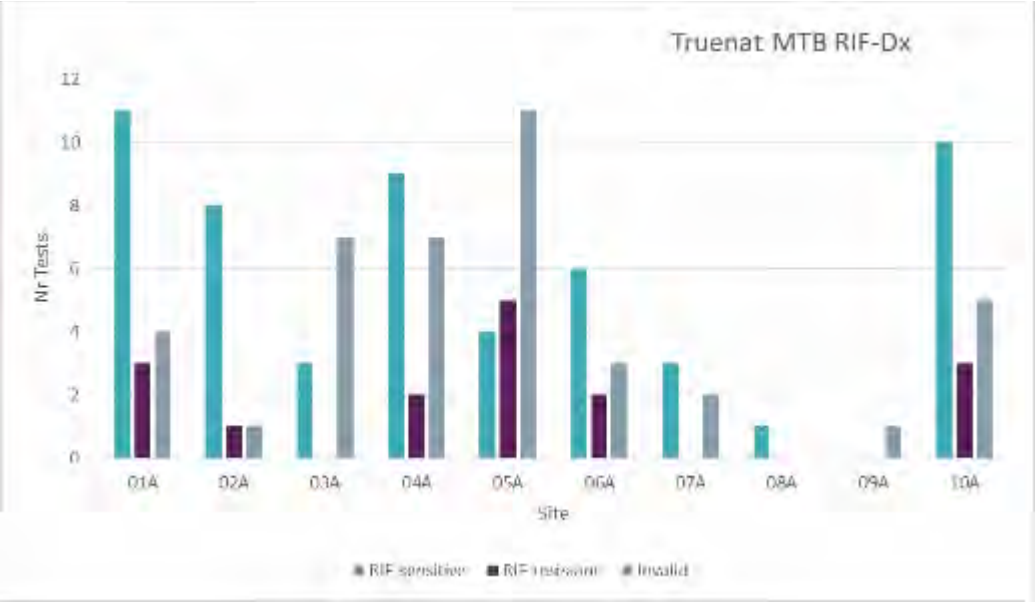


Figure 13: Truenat MTB-RIF Dx testing (among Truenat MTB Plus positives) distribution by site

The overall invalid rates, including those for the DNA extraction procedure (Trueprep), are shown in Table 18. The invalid rates for Truenat MTB Plus and MTB-RIF Dx were higher in the current study compared to a prior study conducted by FIND on frozen specimens (13.3% vs. 2.5% for MTB Plus and 36.6% vs. 33.9% for MTB-RIF Dx). Root-cause analysis identified several steps to improve the assay, which were implemented before the clinical evaluation study (Section 2) was initiated. In brief these include:
i) Removal of Mg2+ from eluate buffer and placing as a component of the microtube with lyophilized primers, enzyme and dNTPs – this increased overall stability of the DNA, which was particularly important for the RIF assay which is a reflex test
(ii) improved stability at of the assay components at higher ambiet and operational temperatures and
(iii) provision of a dedicated fixed-volume (6µL)precision pipette, ensuring consistency in DNA mixing and loading."

Table 20: Invalid/Error rate among sputum specimens processed

Procedure/Assay	Frequency
Trueprep	19.8% (96/486) [1]
Truenat MTB Plus	13.3% (59/443) [2]
RIF-Dx	36.6% (41/112) [3]

[1]Invalid/Error (72/96), no DNA eluate (5/96), cartridge and/or valve error including leakage or clogging (19/96).
[2]Invalid/Error (59/59). [3]Indeterminate/Error (41/41)

Swab testing
In order to monitor potential DNA/amplicon contamination, all sites were instructed to perform swab testing of working areas and of device surfaces at study start, and once a week thereafter. Swabs were tested as follows:
Swab A: specimen preparation area
Swab B: area around the Trueprep device
Swab C: area around the Truelab device
Swab D: cartridge holder inside the Trueprep device
Swab E: chip tray inside the Truelab device

A total of 231 swab tests were done, of which 23 (10%) were positive by Truenat MTB Plus. The results per swab and site are shown in Table 21. Positive swabs were observed at all sites except for 2, of which one was the only reference lab participating in the study.

Table 21: Swab testing results per site

	Swab A		Swab B		Swab C		Swab D		Swab E	
Site	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
01	0	6	1	4	3	3	1	5	1	5
02	1	3	0	5	0	4	0	4	1	4
03	0	2	1	4	0	2	0	4	0	3
04	0	3	0	3	2	1	1	2	1	3
05	1	6	0	7	0	6	1	5	0	7
06	1	2	0	3	0	3	0	3	0	3
07	1	5	0	4	1	5	0	4	3	4
08	0	8	2	4	0	5	0	6	0	8
09	0	5	0	5	0	4	0	6	0	5
10	0	3	0	3	0	3	0	3	0	3
Total	4	43	4	42	6	36	3	42	6	45

Testing of external controls

One positive (PC) and one negative (NC) external control provided by Molbio (Truenat Universal Control Kit) were run at least twice upon setup and at study end. According to manufacturer instructions, 6 µL from either control are transferred onto the Truenat chip and serves to validate the performance of both the chip and the Truelab analyzer.

A total of 96 control runs (tests) were performed throughout the study with a 10.4% invalid rate (10/96). Among the valid runs, the results of all PC were as expected, i.e. positive. However, at 2 sites (site 03 and 08) 3 NC out of 96 were found to be positive by Truenat MTB Plus (2 positive-NC corresponded to the same site, for both the first and repeat run). It was reported by the lab supervisors that the issue occurred during preparation of the PC. As per package insert, the dried-down positive control needs to be reconstituted by adding 50 µL of NC. The controls were repeated using a new kit afterwards and the NC were negative.

Additionally, in order to process external controls, the operator has to select a specific profile from a drop-down menu (under "Sample Type", either "POS control" or "NEG control"). It was observed during the study that operators would often select "Sputum" instead of "POS control" or "NEG control". Therefore, given the way external controls were analysed by the system software for the intended use, i.e. to validate the amplification step, the system would interpret the controls as "Sputum", applying a different QC threshold for passing the assay, than would be applied for an actual clinical specimen. This incorrect profile selection error by the operator therefore resulted in a larger than anticipated proportion of "invalid" result for control testing.

Table 22: Lab technician's appraisal at study end (Questionnaire: User set 3)

Aspect	Appraisal
Biggest benefits of the assay*	*"Early RIF results (and MDR)"** *"May replace smear microscopy"** *"Easy to detect TB patients much early"** *"Requires less time than smear microscopy"**
Biggest disadvantages*	*"High number of invalid results"** *"Only up to 8-10 tests per day (single test at a time)"** *"Requires dedicated (skilled) person… difficult with current lab workload"* *"Difficult to load samples in the chip (silly mistakes yield to invalid result)"** *"Long test running time"**

Aspect	Appraisal
	"Issues with liquefaction (of viscous specimens) and cartridge getting clogged"
How could Truenat be further improved*	*"It should be fully automated like CBNAAT (one-step method)"** *"It should be possible to run multiple specimens simultaneously (not suitable for high workload labs)"** *"Invalid results take 30 minutes to be displayed… time wasted"* *"Running time should be shorter"* *"Specimen processing procedure should be improved (alternative liquefaction buffer)"*
Additional comments*	*"Overall good/easy"** *"Several chips of incorrect size, would not fit in the chip slot"*

*Indicates comments mentioned more than once by the users

The laboratory supervisors were also asked to complete a questionnaire at the end of training and the same questions were also asked at study end. Overall, the impressions at the end of training were maintained and in general almost all considered the i) setup and training, ii) ease of use of devices and accessories and iii) troubleshooting steps to be easy (Table A5).

Conclusions and next steps

Overall, the feedback from laboratory technicians at the microscopy center level, who had no experience on molecular testing, was positive. This was also true of laboratory supervisors involved in the study. Nevertheless, some concerns were raised in terms of the total duration of the assay run, the complexity of the assay (precision pipetting step) and low throughput. Therefore, considerations for easier transfer of DNA eluate would be beneficial. Molbio is looking at providing a fixed-volume (6µL) pipette in the future (similar to a Pasteur transfer pipette) instead of a micropipette.
Further observations by the manufacturer have shown that reduced visibility due to lack of adequate lighting at microscopy centres may contribute to the challenges during the precision pipetting step, thus is looking at providing a USB light attached to the Truelab device.

As for the assay throughput, two new models of Truelab are available since February 2018: the Truelab Duo and the Truelab Quattro with 2 and 4 chip slots, respectively. The invalid rates for Truenat MTB Plus and MTB-RIF Dx were higher than those seen during a prior study on frozen specimens. The removal of Mg2+ from the eluate buffer has subsequently improved DNA stability. Technical issues reported during our study included: Truelab analyzers damaged during transit, Trueprep cartridge holder failure and issues with the chips (loading and interpretation). Before the initiation of the clinical evaluation study, Molbio has provided improved packaging for shipment of equipment and consumables, and improved packaging and marking of IP contents. A new single sterile sleeve for each Truenat chip (containing the chip, the microtube and a sterile pipette tip) has been included within the chip pouch itself, which also contains a desiccant.

Given the number of positive swabs observed during our study, there is a concern for possible DNA cross-contamination. Careful assessment of specificity during longer-term use in microscopy centers is critical. Errors during testing of external controls indicate that careful instructions and training on how to run the external controls correctly is key. Moreover, these controls only serve to validate the amplification step (not the DNA extraction step) and cannot be blinded (and thus not be used for external quality assurance). Therefore, additional controls that could be used for proficiency testing (including DNA extraction and amplification), as well as for external quality assurance, will be required.

Key recommendations for future training and implementation of Truenat:

- Careful consideration of critical steps such as precision pipetting and training on how to run the external controls;
- Close monitoring of invalid results which would have an impact on time, costs and reliability by the end-users;
- Regular testing of negative controls to detect any potential carry-over contamination in microscopy centres where most lab technicians have no experience performing molecular tests;
- Other factors that require careful consideration during implementation include:
    - Capacity of scale-up for setup and adequate training to be done directly by the manufacturer;
    - Microscopy centres should have a sufficient supply of electricity per day or night in order to recharge the equipment. Alternatively, Molbio may want to consider provision of an optional device to allow solar charging;
    - Microscopy centres with no capacity for extended storage of reagents in hot regions (stability is assured for up to 2 years if stored below 30°C, and for up to 6 months if stored below 40°C).
    - The assessment of connectivity features of the Truelab analyzer was not possible during our study. Evaluation in the hands of lab technicians at the microscopy centre level will be required.

Acknowledgements
Laboratory technicians and supervisors participating in the study, for their valuable feedback
NIRT: Mary Rebecca, Nallathambi S, Suganya Devakumari
FIND India: Sanjay Sarin, Loganathan Prabakaran, Aradhana Chauhan
FIND South Africa: André Trollip

This report has been prepared for the WHO Global TB Programme. Do not distribute further.

CHAI: Sachin Salim, Rattan Ichhpujani, Lakshmi Nair
Others: Sophia Georghiou, David Dolinger
Rutgers: David Alland, Yuan Cao, Heta Parmar

Web Annex D.8. Moderate complexity automated NAATs: Diagnostic accuracy for TB detection and detection of resistance to rifampicin and isoniazid. A systematic review and meta-analysis

December 2020

Authors:
Mikashmi Kohli[1,2]; Emily MacLean[1,2]; Madhukar Pai[1,2]; Claudia M. Denkinger[3,4]; Samuel G. Schumacher[3]

Corresponding author:
Samuel Schumacher PhD
Evidence & Policy Lead at FIND
Geneva, Switzerland
Email: Samuel.Schumacher@finddx.org

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada.
2 McGill International TB Centre, McGill University, Montreal, Canada.
3 Foundations for Innovative New Diagnostics, Geneva, Switzerland
4 Center for Infectious Diseases, University Heidelberg, Heidelberg, Germany

EXECUTIVE SUMMARY

Background

The World Health Organization (WHO) estimates that in 2017, 10 million people became ill with TB globally. Drug-resistant TB is an enormous threat. In 2017, an estimated 558,000 people were newly diagnosed with rifampicin-resistant TB, 468,720 of whom had multidrug-resistant TB (MDR-TB). MDR-TB is caused by infection with *M. tuberculosis* bacteria that are resistant to at least rifampicin and isoniazid. The introduction and rollout of nucleic acid amplification tests (NAATs) has revolutionized the area of TB diagnosis. WHO recommends Xpert, LPA, LAMP and Mol Bio TrueNaat assay as the molecular tests to be used for diagnosis of TB and drug resistance. Other diagnostic companies have recently entered the realm of molecular testing for TB and drug resistance detection. These are end-to-end solutions which have been developed for various diseases such as TB, HIV, HPV etc. Some have the capacity to provide results for more than 1000 specimens for TB detection and/or drug resistance in less than 10 hours. Data for this review consists of evidence evaluating seven such molecular assays for TB detection and drug resistance.

Methods

A comprehensive search of the following databases (PubMed, EMBASE, BIOSIS, Web of Science, LILACS, Cochrane) for relevant citations was performed. The search was restricted to the time period January 2009 to July 2020. Reference lists from included studies were also searched. No language restriction was applied. As the number of studies for the index tests are few, we contacted the diagnostic companies for reports of their internal validation data. Studies were also included from the WHO public call to submit the data. The quality of studies was assessed using an adapted QUADAS-2 tool. A culture-based reference standard was used for the evaluation of *Mtb* detection. Resistance detection was compared against a phenotypic reference standard, as well as a composite reference standard (constructed by combining the results of phenotypic and genotypic DST results in studies where both had been performed). Bivariate random-effects meta-analyses were performed using STATA to obtain pooled sensitivity and specificity estimates with 95% confidence intervals (CI) for RIF resistance, INH resistance and *Mtb* detection. Where only a limited number of studies were available, descriptive analyses were conducted.

Results

A total of 36 studies contributed to 39 unique datasets (four studies provided data for more than one index test). All the studies were performed in central laboratories as the index tests require sophisticated laboratory infrastructure. For most studies, samples were included in the study on arrival into the laboratory, as a result there was limited data on the demographics of the included patient population, such as age, HIV status, past TB history.

A total of 29 studies with 13852 specimens (4767 TB positive specimens) provided data for evaluating TB detection from the five index tests. The pooled sensitivity (95% CI) was 93.0% (90.9 to 94.7) and the pooled specificity was 97.7% (95.6 to 98.8). The pooled sensitivity in smear negative specimens was 86.1% (73.4 to 93.2). The sensitivity in smear-positive specimens was 98.8 % (93.7 to 99.7) .
For resistance detection, 18 studies, 2874 specimens (702 rifampicin resistant specimens; 854 isoniazid resistant specimens) provided data for resistance testing of two first line drugs using these centralised platforms.
For rifampicin, the pooled sensitivity (95% CI) was 96.7% (93.1 to 98.4) and the pooled specificity was 98.9% (97.5 to 99.5). For isoniazid, the pooled sensitivity (95% CI) was 86.4% (82.8 to 89.3) and the pooled specificity was 99.2% (98.1 to 99.7). These estimates were with phenotypic DST as the reference standard. With composite reference standard, the estimates were very similar.
Data was insufficient to do any analysis on other covariates like HIV status, percentage of children in the study, and treatment status.

Conclusion

In patients with pulmonary TB, the centralized molecular assays demonstrate promising diagnostic accuracy for TB detection, RIF resistance and INH resistance. The diagnostic accuracy estimates of this class of technology appears in comparable range of diagnostic accuracy to the WHO recommended molecular tests for pulmonary TB detection and resistance detection.

Background

Tuberculosis (TB) is caused by *Mycobacterium tuberculosis (Mtb)* bacteria. TB causes tremendous suffering worldwide and has surpassed HIV/AIDS as the world's leading infectious cause of death. The World Health Organization (WHO) estimates that in 2017, 10 million people became ill with TB globally. Approximately, 1.3 million HIV-negative people and 300,000 HIV-positive people died from TB (WHO 2018). Drug-resistant TB is an enormous threat. In 2017, an estimated 558,000 people were newly diagnosed with rifampicin-resistant TB, 468,720 of whom had multidrug-resistant TB (MDR-TB). MDR-TB is caused by infection with *Mtb* bacteria that are resistant to at least rifampicin (RIF) and isoniazid (INH).

The introduction and rollout of nucleic acid amplification tests (NAATs) has revolutionized the area of TB diagnosis by providing rapid and accurate diagnostics (WHO 2010). The principal behind these tests is amplification of the targeted genomic region of the *Mtb* bacteria using polymerase chain reaction (PCR). NAATs are used for both TB detection and analysis of anti-TB drug resistance, most commonly RIF and INH (UNITAID 2017; Dicks 2019).

Resistance to RIF is usually associated with mutations in the hot-spot region of the *rpoB* gene. Resistance to INH is observed due to mutations in many genes. Mutations in *katG* and *inhA* genes are the most frequently observed and have been the bases of many molecular diagnostic tests for detecting INH resistance. However, there are other rarer mutations in other genes like *ahpC*, *fabG1*, *kasA,* and *efpA* (Almedia 2011; Somoskovi 2001). Recently, WHO made evidence-based recommendations for INH mono-resistant TB cases and suggested a non-standard treatment for these cases. Globally, INH monoresistance is more prevalent than MDR-TB and these guidelines advocates for universal testing for both RIF and INH resistance at the start of TB treatment (WHO 2018).

Thus far the molecular market has been dominated by Cepheid (US) with the GeneXpert, a near patient platform that allowed for *Mtb* and RIF resistance detection. Bruker-Hain (Germany) has two line-probe assays available for RIF/INH detection (GenoType MTBDR*plus*) and FQ/SLID detection (GenoType MTBDRsl) for centralized settings. Recently, WHO also recommended LAMP and MolBio True NAAT for TB and drug resistance detection (WHO, 2019)

Recently, several companies have developed molecular tests for TB and RIF/INH resistance detection on centralized platforms, many of which have been established as multi-disease platforms, primarily for detection of Human immunodeficiency virus (HIV), Human Papillomavirus (HPV) and Hepatitis C Virus. On these high-throughput platforms a large number of patient specimens can be tested in a shorter amount of time. Some assays can generate around 1000 results in just 8 hours (Cobas 8800 platform). This could help reduce the current problem of diagnostic delays and delay in treatment initiation for TB patients (Subbaraman 2016), if transport logistics and communication of test results can be ensured.

This systematic review intends to evaluate the data available on the diagnostic test accuracy of these tests for *Mtb* and RIF/INH resistance detection.

Index tests included in this systematic review

Abbott Molecular was the first company to develop two centralized NAATs, one for TB detection (RealTi*m*e MTB assay) and one for RIF/INH (RealTi*m*e MTB RIF/INH). The Abbott RealTi*m*e MTB RIF/INH resistance assay uses eight dye-labeled probes to detect the RIF resistance determining region of *rpoB* gene and four probes for IHN, with two probes each for *katG* and *inhA* genes. The limit of detection (LoD) has been reported as 17 cfu/mL for the RealTi*m*e MTB assay and as 60 cfu/mL for the RealTi*m*e RIF/INH assay (UNITAID 2017; Abbott 2019a; Abbott 2019b). The assays operate on the high throughput m2000 platform, m2000*sp* for fully automatic DNA extraction and m2000*rt* for performing the real time PCR.

Bruker-Hain Diagnostics has expanded its portfolio from the line-probe assays to offer the FluoroType® MTB to detect *Mtb* DNA and FluoroType® MTBDR, which can detect RIF and INH resistance in addition. These are completely independent platforms and have no relation to the GenoType MTBDR platforms. Both assays are CE- marked since 2014 and 2018, respectively. The assays utilize asymmetric excess PCR and light on/off probes. The target genes that are detected in this assay are *rpoB* for RIF and *inhA* promoter and *katG* gene for INH resistance. The LoD of FluoroType MTB assay has been reported as 15 cfu/mL and as 20 cfu/mL for FluoroType MTBDR assay (UNITAID 2017; Bruker-Hain 2019a; Bruker-Hain 2019b). For DNA extraction, there are manual (FluoroLyse) and automated (GenoXtract) options available. The platform used for amplification and detection are FluoroCycler® and FluoroCycler® XT for MTB and MTBDR assays respectively. Both have high throughput.

Becton Dickinson (BD) has developed a multiplexed real-time PCR (BD MAX™ MDR-TB, CE-marked since 2018) assay for the detection of *Mtb* and resistance to both RIF and INH. The platform utilizes 5-colour detection (UNITAID 2017). For *Mtb* detection, this assay detects multicopy genomic targets IS6110 and IS1081 as well as a single copy genomic target. For resistance, the assay detects *rpoB*, *inhA* promoter and *katG* genes. The LoD of BD Max MDR-TB assay has been reported as 0.5 cfu/ml for *Mtb* detection and 6 cfu/ml for resistance detection. The assay uses a fully-integrated, automatic, high-throughput BD Max™ platform.

Roche Diagnostics (Roche) has developed COBAS MTB and MDR-TB assay to detect TB and drug resistance on their COBAS® 6800/8800 Systems, which are fully automated and very high throughput (e.g. the 8800 system can process 960 tests in an 8-hour period) (UNITAID 2017). The assay detects both 16S rRNA and *esx* genes as target genes for *Mtb* detection. The LoD for this assay ranged from 7.6 cfu/mL to 8.8 cfu/mL.

Further details on the tests and platforms is available in a separate analysis of operational aspects conducted by FIND.

Methods

We followed standard guidelines and methods for systematic review and meta-analyses of diagnostic test accuracy (Moher 2009; Cochrane 2008). We prepared a protocol for the literature search, article selection, data extraction, assessment of methodological quality and synthesis of results.

Selection criteria

Types of studies (designs)

Cross-sectional, case-control, cohort studies or randomized controlled trials comparing any of the above-mentioned molecular tests to a reference standard test (see section below on Reference Standard) were included, if at least 25 specimens were tested. Given the limited number of studies evaluating the new molecular diagnostic assays to date, case-control studies were included, as long as cases and controls were sampled from the same patient population. We also included studies that were submitted though the WHO public call on these platforms.

Types of participants

Patients of all age groups presumed of having or confirmed with pulmonary TB or MDR-TB, in all settings and all countries, were included. Specimen types were limited to pulmonary TB specimens (such as sputum, bronchoalveolar lavage, tracheal aspirate etc.).

Index tests

1. Abbott RealTime MTB and Abbott RealTime MTB RIF/INH
2. Bruker-Hain FluoroType MTB and FluoroType MTBDR
3. BD Max MDR-TB assay
4. Roche Cobas MTB assay and Roche Cobas MTB RIF/INH assay

Target conditions

We considered the following three target conditions: resistance of *Mtb* to rifampicin; resistance of *Mtb* to isoniazid; and disease caused by pulmonary *Mtb*.

Reference standard

The reference standard for the detection of *Mtb* was a positive culture for *Mtb*. The specific culture method and number of cultures inoculated were recorded: Liquid (MGIT 960, BACTEC 360), Solid (LJ, 7H10, 7H11), Mixed (i.e. more than one culture method was used) or "Other".

The reference standard for the detection of RIF and INH resistance detection was primarily culture-based, i.e. drug susceptibility testing (DST) as the primary reference standard for resistance detection. In addition, results from genetic sequencing were obtained (genotypic DST), where reported and a composite reference standard was developed, which combined the results from genotypic and phenotypic DST results.

For composite reference standard, if conventional DST showed sensitivity but sequencing identified mutations recognized to be associated with resistance, the composite reference standard was considered resistant if the mutations were associated with high or moderate confidence of being associated with resistance as per Miotto et al. (Miotto 2017). If conventional DST showed resistance but sequencing did not identify mutations to be associated with resistance, the composite reference standard was considered resistant (as mutations will be assumed outside of the region sequenced or alternatively there may be

low-level heteroresistance, below the limit of detection of the sequencing technologies used; e.g. Sanger). The composite reference standard was constructed without knowledge of the index test results.

Outcome measures

For *Mtb* detection, sensitivity refers to the proportion of specimens or isolates with a culture positive for *Mtb* that were detected as *Mtb*-positive by the index test.
For RIF- or INH-resistance detection, sensitivity refers to the proportion of specimens or isolates with the phenotypic or composite reference standard separately, demonstrating RIF/INH-resistance that were detected as RIF/INH-resistant by the index test.

For *Mtb* detection, specificity refers to the proportion of specimens or isolates with a culture negative for *Mtb* that were detected as *Mtb*-negative by the index test.
For RIF- or INH-resistance detection, specificity refers to the proportion of specimens or isolates with the phenotypic or composite reference standard separately, demonstrating RIF/INH-sensitivity that were detected as RIF/INH-sensitive by the index test.

Search methods

A comprehensive search of the following databases (PubMed, EMBASE, BIOSIS, Web of Science, LILACS, Cochrane) for relevant citations was performed (full search strategy reported in Appendix A). The search was restricted to the time period January 2009 to July 2020. Reference lists from included studies were also searched.

No language restriction was applied. As the number of studies for these index tests are less, we included abstracts which were not published as a full research paper. Various diagnostic companies were also contacted for their internal validation data and data submitted through the WHO public call was also included.

Study selection methods

Two review authors (MK and EM) independently assessed titles and abstracts identified by electronic literature searching to identify potentially eligible studies (Screen 1). Any citation identified by either review author during screen 1 was selected for full-text review. The same two review authors (MK and EM) independently assessed the full-text articles for inclusion using predefined inclusion and exclusion

criteria (Screen 2). In Screen 2, any discrepancies were resolved by discussion between the review authors, and for any study that were not resolved by the review authors, were then decided by a third review author (SS). If a study contributed data to more than one analysis (e.g. two different index tests in one study, on the same or different specimens), it was considered as two or more datasets.

Data extraction

We created a data extraction form, piloted it with a subset of eligible studies, and then finalized the form (Appendix B). Two review authors (MK and EM) independently extracted data from the included studies with the standardized form and crosschecked to ensure accuracy. Disagreement between review authors on data extraction was resolved by discussion or by a third reviewer (SS). For studies without complete extraction information available, authors were contacted to request further data. Studies without extractable sensitivity and specificity data were excluded if no further information was acquired after three attempts to contact the study authors.

Assessment of methodological quality

The QUADAS-2 instrument, a validated tool for diagnostic studies, was used to assess study quality (Whiting 2011). The information needed to answer QUADAS-2 questions was incorporated in the data extraction sheet. A description of the QUADAS-2 items and the interpretation in the study context can be found in Appendices C.

Statistical analysis and data synthesis

Statistical analyses were performed using Stata (version 14; StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP). The studies were grouped by type of index test and reference standard used. QUADAS-2 analysis was performed using Excel (version 14.5.4; Microsoft, Seattle, WA).

Approach to indeterminate index test results

Indeterminate test results were excluded from the analyses for determination of sensitivity and specificity and were reported separately.

Assessment of publication bias

Formal assessment of publication bias (tests for funnel plot asymmetry) was not performed, as these techniques are not recommended for diagnostic test accuracy studies (Cochrane 2008). Due to very little data available on this topic, unpublished data were also included in this review.

Meta-analysis

Meta-analysis was after pooling the dataset from all the index tests together. We also visually saw the results from each index test by forest plots. Bivariate random effects meta-analyses were performed (Chu 2006; Reitsma 2005) using the metandi programme in Stata for index tests with enough studies that included data to calculate sensitivity, specificity, and 95% confidence intervals for each of these. Summary and individual estimates were also presented graphically with the 95% Confidence Interval (CI) interval and prediction region.

Forest plots were visually assessed for heterogeneity among the studies within each index test and in the summary plots we examined the variability in estimates and the width of the prediction region, with a wider prediction region suggesting more heterogeneity.

There was a pooling criteria that was developed and consensus was reached to use that pooling criteria by inputs from WHO and the GDG members. The pooling criteria to pool the data from different index tests is given in the table below:

Parameters	**Sensitivity**	**Specificity**
Pre-conditions	**n≥50 TB+ (number resistant for resistance detection)**	**n≥100 TB- (number susceptible for resistance detection)**
Condition 1	The pooled estimate of one assay lies within +/-5% of the overall point estimate	The pooled estimate of one assay lies within +/-2% of the overall point estimate
Condition 2	The pooled estimate for one assay lies within 95%CI of the overall point estimate AND The pooled estimate for one assay lies within +/-10% of the overall point estimate	The pooled estimate for one assay lies within 95%CI of the overall point estimate AND The pooled estimate for one assay lies within +/-5% of the overall pooled estimate

Assessment of the quality of evidence

We assessed the quality of evidence using the GRADE approach (Schünemann 2020; Schünemann 2020). We used the GRADEpro Guideline Development Tool software (GRADEpro 2015) to generate

Summary of Findings tables. As recommended, we rated the quality of evidence as high (no points subtracted), moderate (one point subtracted), low (two points subtracted), or very low greater than two points subtracted) based on five factors: risk of bias, indirectness, inconsistency, imprecision, and other considerations. We subtracted one point when there was a serious issue identified or two points when there was a very serious issue identified in any of the factors used to judge the quality of evidence. Two review authors (MK and EM) discussed and applied the GRADE criteria.

Results

Results of the search

From the literature search, 878 citations were identified, 133 full-text articles were reviewed: 36 studies were included in this systematic review (see Figure 1). These 36 studies contributed 39 datasets (3 provided data for more than one index test). All studies were conducted in central level laboratories, which was expected as these centralized assays require sophisticated laboratory infrastructure and skilled laboratory workers. Thirteen studies (36%) utilized specimens from high-income countries or were conducted in high-income countries; 14 studies (39%) utilized specimens from middle-income countries or were conducted in middle-income countries; three studies (8%) evaluated specimens from middle and high-income countries, one study (3%) combined specimens from low- and high-income countries and five studies (14%) utilized specimens from low and middle-income countries (LMICs) or were conducted in LMICs.

There were five assays which contributed data for TB detection. Abbott Realtime MTB, FluoroType MTB, FluoroType MTBDR, BD Max MDR-TB and Cobas 6800/8800 MTB. Of these, all but Cobas MTB assay (n = 2) had sufficient data to be meta-analysed.

The following sections provide the results of the systematic review and meta-analysis for all the index tests. The methodological quality for all the included studies were assessed using QUADAS-2 tool (Figures 2 and Figure 7).

TB Detection

1. PICO question: Among people with signs and symptoms of TB (adults, PLHIV and children) should cNAT assays on sputum be used as an initial diagnostic test for PTB?

A total of 29 studies with 13852 specimens provided data for evaluating TB detection from the five index tests. There were 12 studies that were performed on the Abbott RealTime MTB assay, six on FluoroType MTB, four on FluoroType MTBDR, five on BD Max and two on Cobas MTB assay. The reference standard for each of these studies for TB detection was culture based detection.

Methodological Quality assessment

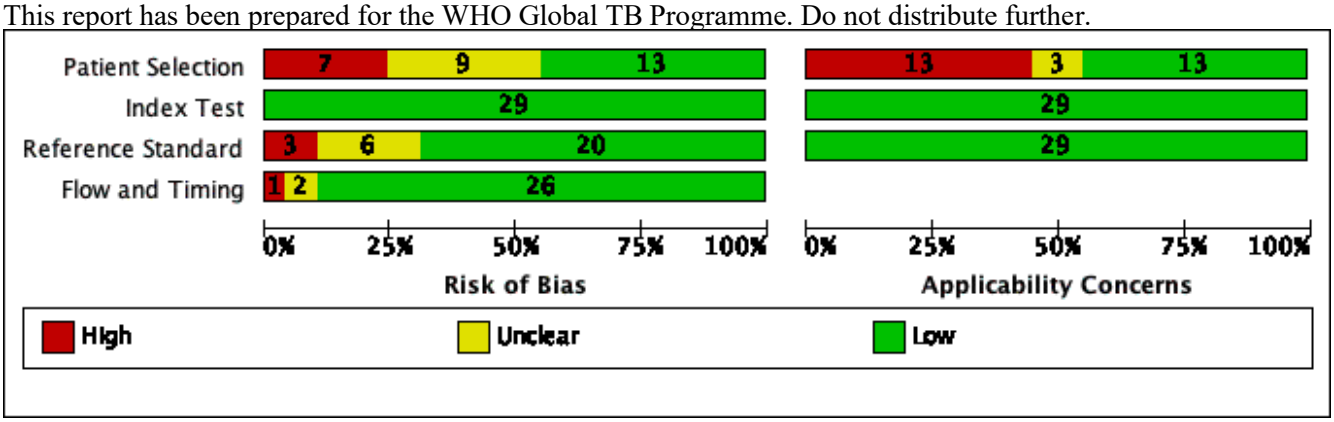


Figure 2: QUADAS-2 assessment for TB detection

Patient selection

Risk of bias

Of the total 29 studies, 16 (55%) had high or unclear risk of bias as they either did prior testing before including specimens in the study or used convenience sampling or the method of participant selection was not reported.

Applicability concern

Only 45% of the studies were conducted in high TB/MDR TB burden country.

Index test

As all the index test assays were automated and thresholds were pre-defined, there is no scope of subjective interpretation of results. We judged all of them at low concern for any risk of bias.

Reference Standard

Of the 29 studies, 9 (31%) had high or unclear risk of bias because either the results of the reference standard were not blinded or it was not reported.

Pooled Analysis

The overall sensitivity in these 29 studies ranged from 79% to 100% and the specificity ranged from 60% to 100%.

The pooled sensitivity (95% CI) was 93.0% (90.9 to 94.7) and the pooled specificity was 97.7% (95.6 to 98.8). Table 2a (Figure 3) provides the overall pooled estimates for all tests and also provides meta-analysed estimates of each index test separately.

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Berhanu 2018	44	6	12	175	0.79 [0.66, 0.88]	0.97 [0.93, 0.99]
Chen 2015	126	2	0	392	1.00 [0.97, 1.00]	0.99 [0.98, 1.00]
Cieslekzuk 2020	36	2	3	28	0.92 [0.79, 0.98]	0.93 [0.78, 0.99]
Cobas 2020	391	6	21	326	0.95 [0.92, 0.97]	0.98 [0.96, 0.99]
de Vos 2018	322	0	13	105	0.96 [0.93, 0.98]	1.00 [0.97, 1.00]
de Vos 2020	231	9	21	196	0.92 [0.88, 0.95]	0.96 [0.92, 0.98]
de Vos 2020-FT	294	17	27	259	0.92 [0.88, 0.94]	0.94 [0.90, 0.96]
de Vos 2020-FTMTBDR	291	7	33	270	0.90 [0.86, 0.93]	0.97 [0.95, 0.99]
Eigner 2013	77	21	4	559	0.95 [0.88, 0.99]	0.96 [0.95, 0.98]
Genc 2018	15	0	1	108	0.94 [0.70, 1.00]	1.00 [0.97, 1.00]
Haasis 2018	64	1	6	271	0.91 [0.82, 0.97]	1.00 [0.98, 1.00]
Hinic 2017	13	3	1	181	0.93 [0.66, 1.00]	0.98 [0.95, 1.00]
Hofmann-Thiel 2020	213	0	33	257	0.87 [0.82, 0.91]	1.00 [0.99, 1.00]
Hofmann-Thiel 2014	45	17	7	902	0.87 [0.74, 0.94]	0.98 [0.97, 0.99]
Hofmann-Thiel 2016	201	2	14	391	0.93 [0.89, 0.96]	0.99 [0.98, 1.00]
Hofmann-Thiel 2016-FT	196	0	19	393	0.91 [0.87, 0.95]	1.00 [0.99, 1.00]
Howlett 2020	96	8	22	299	0.81 [0.73, 0.88]	0.97 [0.95, 0.99]
Luukinen 2020	22	0	0	131	1.00 [0.85, 1.00]	1.00 [0.97, 1.00]
Obasanya 2017	62	91	8	135	0.89 [0.79, 0.95]	0.60 [0.53, 0.66]
Paradis 2018	274	81	13	207	0.95 [0.92, 0.98]	0.72 [0.66, 0.77]
Scott 2017	53	10	9	121	0.85 [0.74, 0.93]	0.92 [0.86, 0.96]
Shah 2019	262	17	20	593	0.93 [0.89, 0.96]	0.97 [0.96, 0.98]
Shaheed 2020	231	7	13	15	0.95 [0.91, 0.97]	0.68 [0.45, 0.86]
Shaheed 2020-FTMTBDR	209	0	38	23	0.85 [0.80, 0.89]	1.00 [0.85, 1.00]
Tam 2017	152	27	9	422	0.94 [0.90, 0.97]	0.94 [0.91, 0.96]
Tang 2015	198	11	14	348	0.93 [0.89, 0.96]	0.97 [0.95, 0.98]
Vinuesa 2016	106	21	5	125	0.95 [0.90, 0.99]	0.86 [0.79, 0.91]
Vinuesa 2018	18	7	0	1384	1.00 [0.81, 1.00]	0.99 [0.99, 1.00]
Wang 2016	159	15	0	81	1.00 [0.98, 1.00]	0.84 [0.76, 0.91]

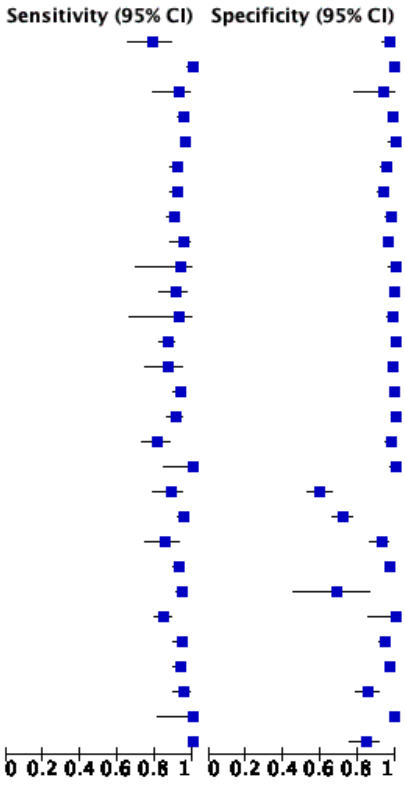


Figure 3: Forest plot of included studies for TB detection

Subgroup analysis: Smear status

There were 26 studies that provided information for TB detection stratified by smear status. A total of 2880 smear positive and 9447 smear negative specimens were evaluated.

The sensitivity for smear positive specimens ranged from 93% to 100%. The pooled sensitivity was 98.8% (93.7 to 99.7) (Table 2b) (Figure 4).

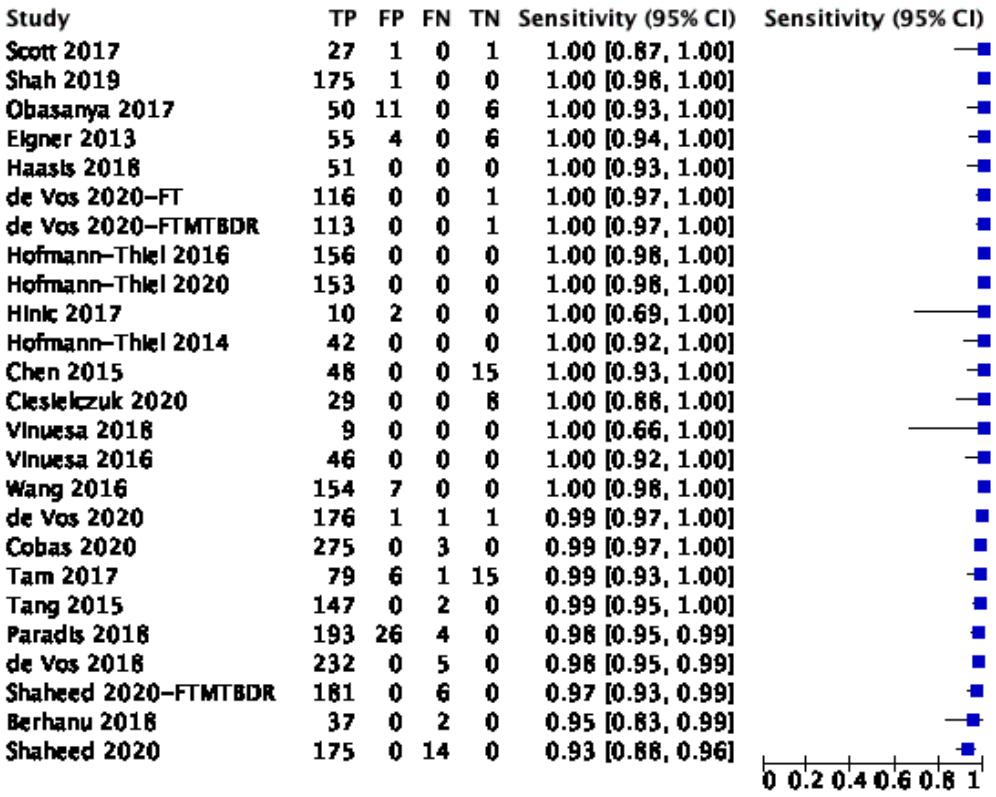


Figure 4: Forest plot of included studies for TB detection in smear positive specimens.

For smear negative the sensitivity ranged from 22% to 100%, and specificity ranged from 68% to 100%. Meta-analysis of the data provided a pooled sensitivity of 86.1% (73.4 to 93.2) and the pooled specificity was 96.4% (93.6 to 98.0) (Table 2b) (Figure 5).

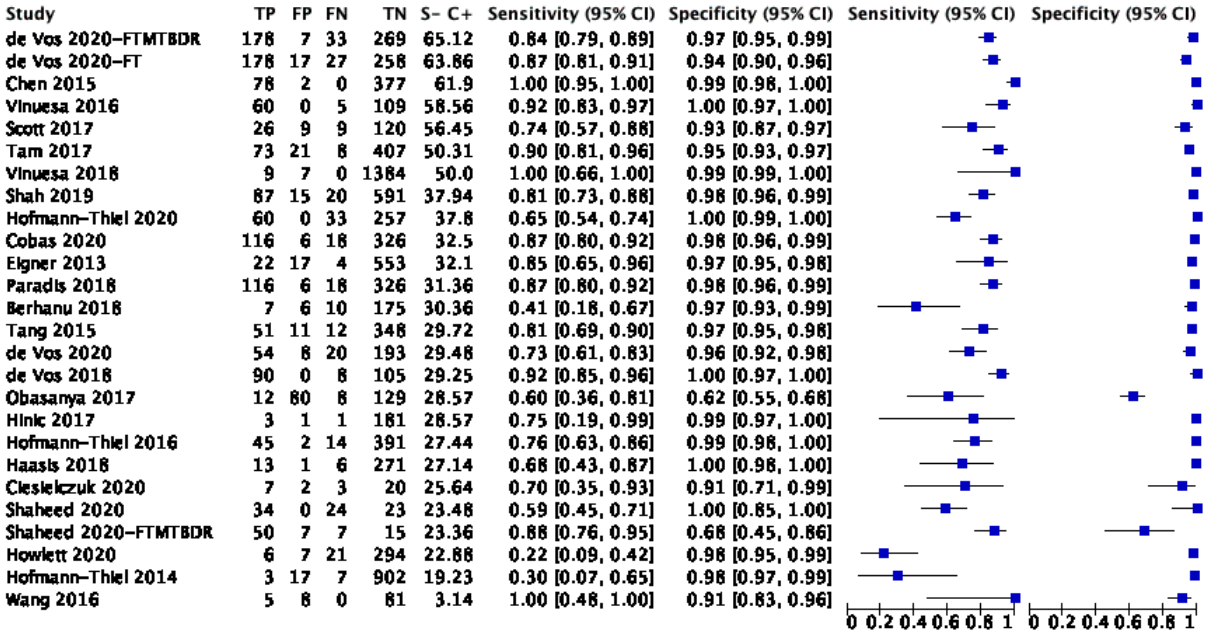


Figure 5: Forest plot of included studies for TB detection in smear negative specimens.

Sensitivity analysis: Manufacturer driven

There were four studies which were manufacturer driven in this dataset for TB detection. We excluded these studies and performed a sensitivity analysis to evaluate the diagnostic accuracy of these class of tests for TB detection. The definition of manufacturer driven was that we removed studies which were provided by the manufacturers' during the public call. Additionally, we reviewed the included published studies and if the authors of any study were manufacturers' and they were involved in the design, analysis or running the study, we also excluded them from our sensitivity analysis.
A total of four studies (Cobas 2020, Hinic 2017, Paradis 2018, and Tang 2015) were excluded from the meta-analysis.

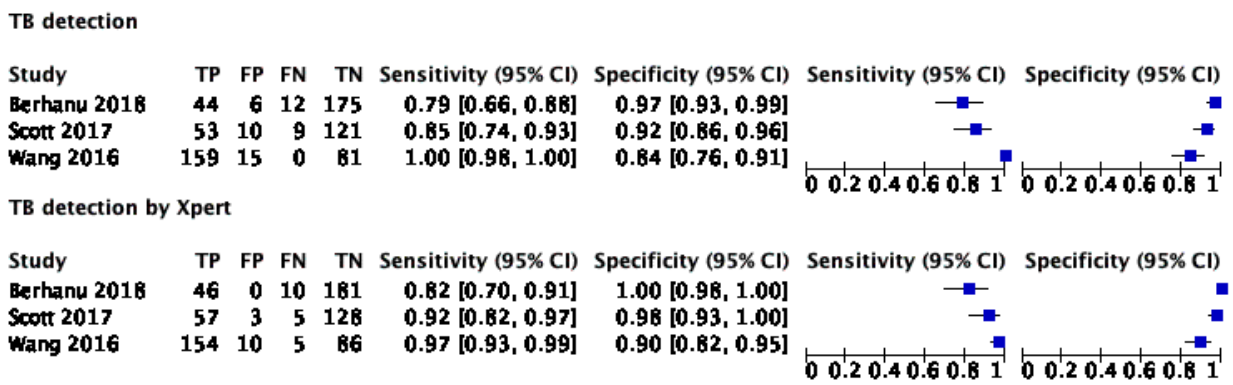
The pooled sensitivity for this analysis was 92.9% (90.4 to 94.8) and the pooled specificity was 98.1% (95.8 to 99.2). Table 4 provides the estimates for sensitivity analyses.

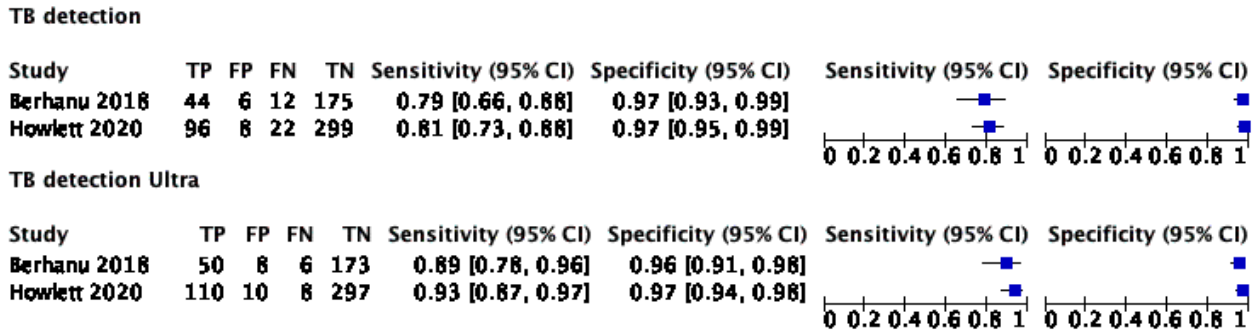
Head-to head comparisons

Whenever the data was available, we did head-to head comparisons of the index test with the well-established and characterised test like Xpert MTB/RIF or Xpert Ultra.
These head-to-head comparisons provide a better understanding of the diagnostic accuracy of the index test as the comparison to well established molecular tests enables a benchmarking in the respective population. It helps understanding if the point estimate for diagnostic accuracy is driven by the test characteristic or it could be attributed to the population characteristics.

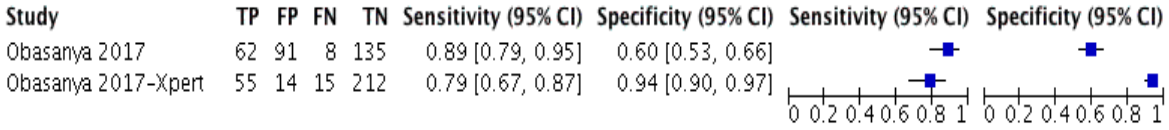
We did not pool the estimates from head to head comparison, as visualizing them separately helps understand whether the point estimates are driven by the test or the study characteristics. We provide comparative forest plots for these comparisons, which help see the results of the index test and Xpert on the same specimens.

Abbott RealTime MTB

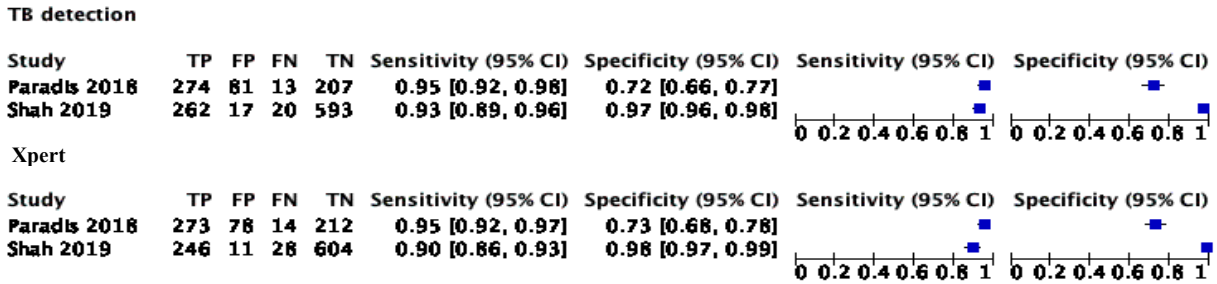




FluoroType MTB



BD Max



Cobas MTB

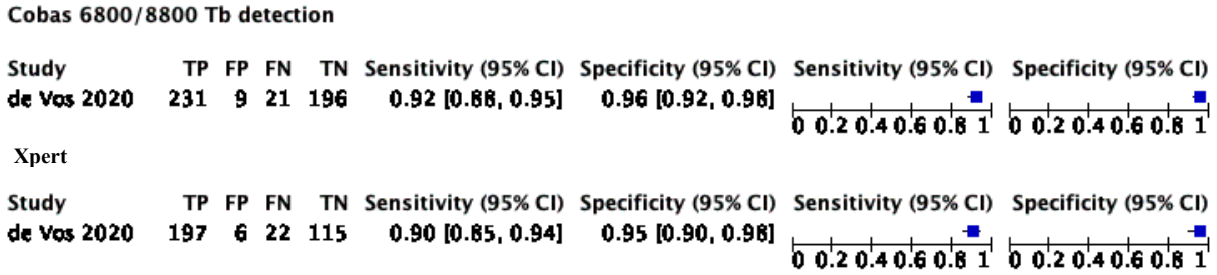


Figure 6: Head- to head comparisons of cNATs and Xpert for TB detection

In the study by Wang 2016, a lower specificity was observed for both Abbott RealTi*m*e MTB assay (84%) and the Xpert assay (90%) for pulmonary TB specimens.
Obasanya 2017 used Xpert in the study for pulmonary specimens. The study reported lower sensitivity of 79% for Xpert in comparison to the FluoroType (sensitivity of 89%), while the specificity of the Xpert

was substantially improved (94% versus 60% for the FluoroType). Considering the possible reasons mentioned about for the low specificity of the FluoroType, and from the head-to-head comparison results, it is possible that the manual DNA extraction method used in this study for performing FluoroType MTB assay could have led to substantially high cross-contamination than a fully automated Xpert assay.
For Paradis 2018, Xpert was also performed on the sputum specimens. It was observed that the specificity was low (73%) with Xpert as well. This suggests that the reasons for the false positives of the index tests, are not specific to the index test but more a problem related to the study.

Forest plot for each index test separately has been provided in Appendix C.

Resistance detection

2. PICO question: Among people with bacteriologically-confirmed PTB (adults, PLHIV, children) should cNAT assays on sputum be used to detect rifampicin resistance using phenotypic DST as reference standard ?

A total of 18 studies, 2874 specimens provided data for resistance testing of two first line drugs (Rifampicin and Isoniazid) using these centralised platforms. There were nine studies that were conducted on Abbott RealTime RIF/INH assay, three on FluoroType MTBDR, four on BD Max and two on Cobas RIF/INH assay. The reference standard for each of these studies for resistance detection was phenotypic DST and a composite reference standard with both phenotypic DST and sequencing results.

Methodological Quality assessment

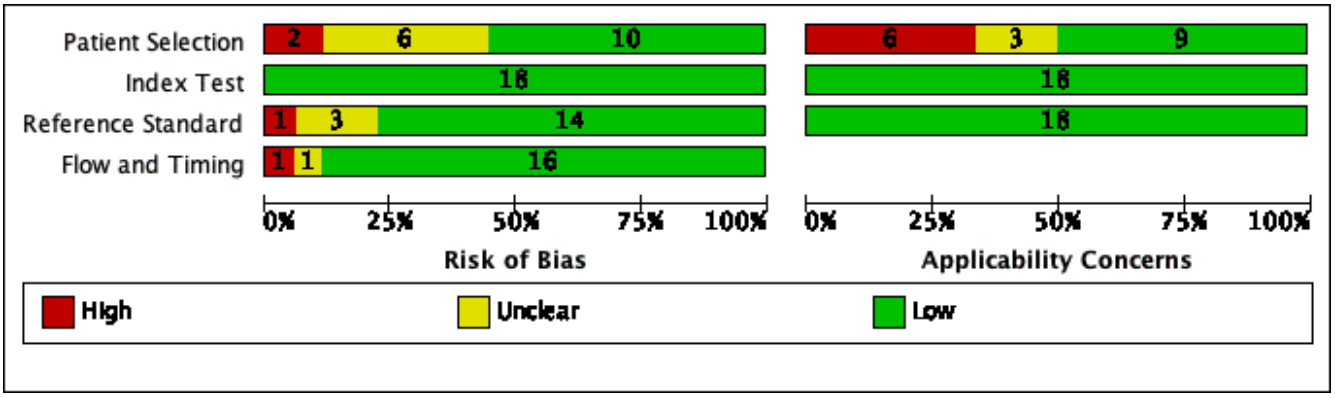


Figure 7: QUADAS-2 assessment for resistance detection

Pooled Analysis: Phenotypic DST

Rifampicin resistance

The overall sensitivity for rifampicin resistance in these 18 studies ranged from 88% to 100% and the specificity ranged from 98% to 100%.
The pooled sensitivity (95% CI) was 96.7 (93.1 to 98.4) and the pooled specificity was 98.9 (97.5 to 99.5). Table 3a (Figure 8) provides the overall pooled estimates for all tests and also provides meta-analysed estimates of each index test separately.

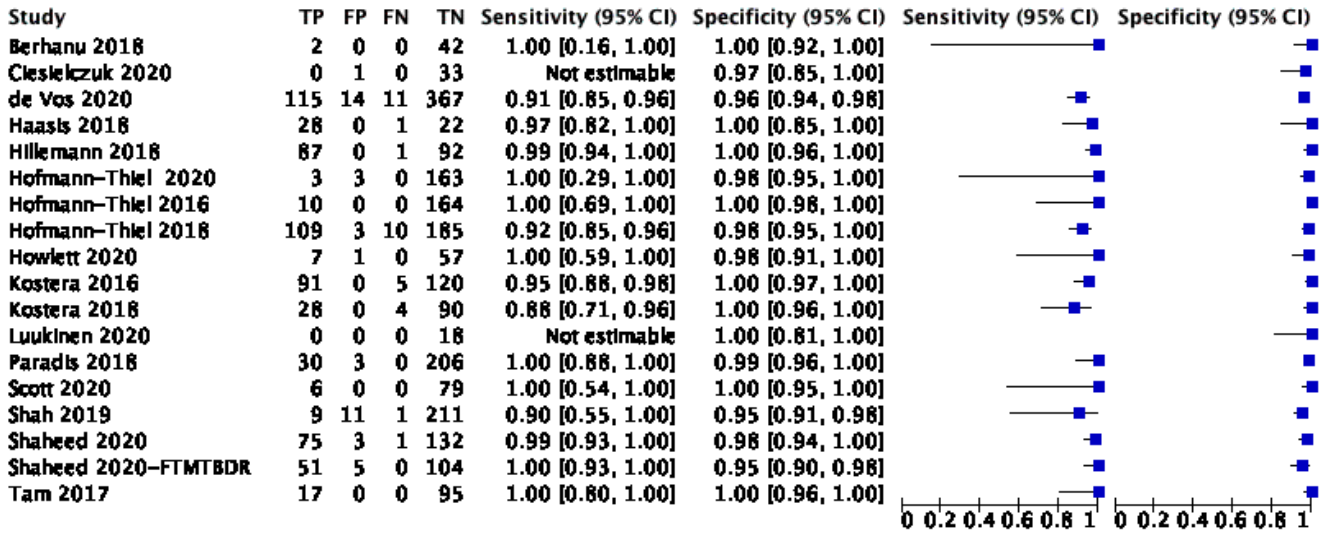


Figure 8: Forest plot for rifampicin resistance detection with phenotypic DST as the reference standard

3. PICO question: Among people with bacteriologically-confirmed PTB (adults, PLHIV, children) should cNAT assays on sputum be used to detect isoniazid resistance using phenotypic DST as reference standard ?

Isoniazid resistance

The overall sensitivity for rifampicin resistance in these 18 studies ranged from 58% to 100% and the specificity ranged from 94% to 100%.
The pooled sensitivity (95% CI) was 86.4 (82.8 to 89.3) and the pooled specificity was 99.2 (98.1 to 99.7). Table 3a (Figure 9) provides the overall pooled estimates for all tests and also provides meta-analysed estimates of each index test separately.

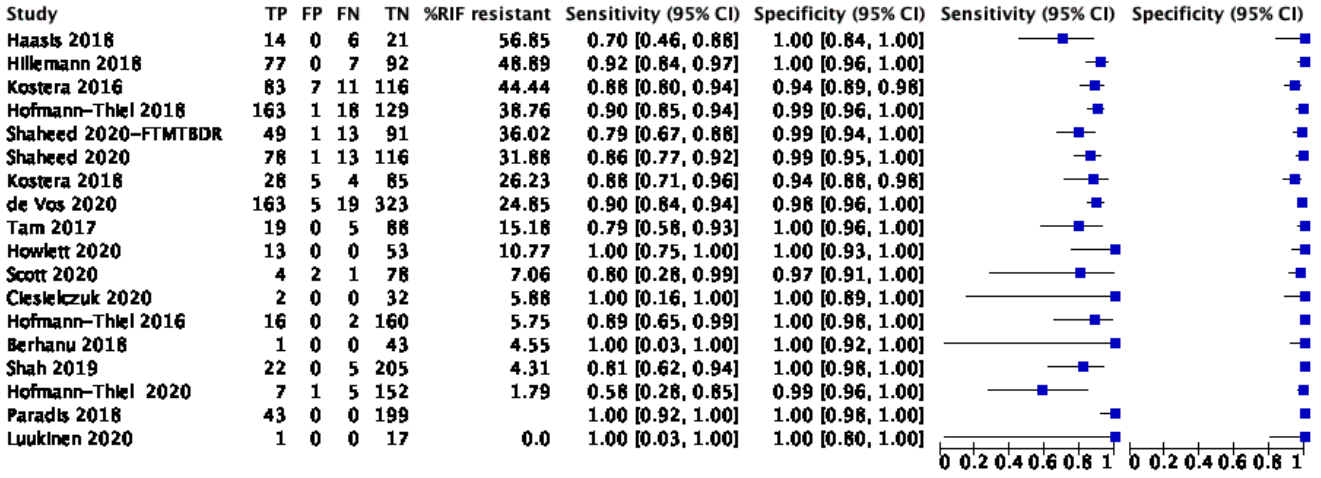


Figure 9: Forest plot for isoniazid resistance detection with phenotypic DST as the reference standard

Pooled Analysis: Composite reference standard

4. PICO question: Among people with bacteriologically-confirmed PTB (adults, PLHIV, children) should cNAT assays on sputum be used to detect rifampicin resistance using composite reference standard ?

Rifampicin resistance

The overall sensitivity for rifampicin resistance in these 9 studies ranged from 92% to 100% and the specificity ranged from 95% to 100%.
The pooled sensitivity (95% CI) was 96.7 (92.9 to 98.5) and the pooled specificity was 98.7 (97.1 to 99.4). Table 3b (Figure 10) provides the overall pooled estimates for all tests and also provides meta-analysed estimates of each index test separately. Details of the discrepant mutations is given in the Appendix D.

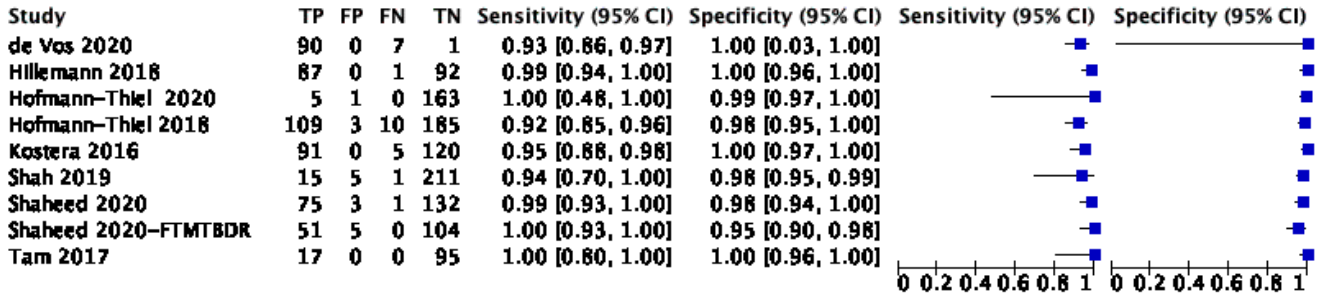


Figure 10: Forest plot for rifampicin resistance detection with composite reference standard

5. PICO question: Among people with bacteriologically-confirmed PTB (adults, PLHIV, children) should cNAT assays on sputum be used to detect isoniazid resistance using composite reference standard ?

Isoniazid resistance

The overall sensitivity for rifampicin resistance in these 8 studies ranged from 79% to 92% and the specificity ranged from 99% to 100%.
The pooled sensitivity (95% CI) was 86.4 (82.1 to 89.8) and the pooled specificity was 99.8 (98.3 to 99.8). Table 3b (Figure 11) provides the overall pooled estimates for all tests and also provides meta-analysed estimates of each index test separately. Details of the discrepant mutations is given in the Appendix C

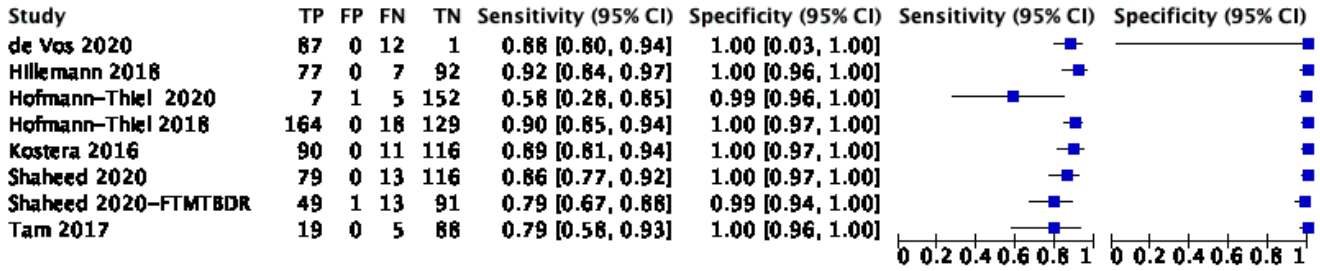


Figure 11: Forest plot for isoniazid resistance detection with composite reference standard

Sensitivity analysis: Manufacturer driven

There were three studies which were manufacturer driven in this dataset for TB detection. We performed a sensitivity analysis excluding these studies driven by manufacturers to assess whether this would affect the results. We excluded Kostera 2016, Kostera 2018 and Paradis 2018.

The pooled estimate for this sensitivity analysis for rifampicin was sensitivity of 97.6% (93.3 to 99.2), specificity of 98.6% (97.0 to 99.4). Similarly, for isoniazid the sensitivity and specificity was 85.7% (81.3 to 89.2) and 99.4% (98.6 to 99.7) respectively. Table 4 gives the sensitivity analysis for resistance detection.

Discussion

This review summarizes the current literature on seven End-to-End solutions for both TB detection and rifampicin and isoniazid resistance detection. Overall, the performance data on the reviewed assays appear promising.

In patients being evaluated for pulmonary TB and resistance detection, these centralized molecular assays demonstrate promising diagnostic accuracy for TB detection, RIF resistance and INH resistance. The performance of these assays is likely similar to that of WHO recommended Xpert and LPA assays. The assays might prove to have operational advantages in some settings due their large throughput.

Acknowledgments

We would like to thank all of the study authors who provided additional data necessary to complete this review. We are grateful to Genevieve Gore, medical librarian at McGill University for assistance with our search strategy. We would also like to thank the manufacturers of these platforms for providing data from their internal company validations.

TABLES

Table 1: Number of included studies for each index test

INDEX TEST	PURPOSE	NUMBER OF STUDIES
ABBOTT REALTIME MTB	TB detection	12
ABBOTT REALTIME RIF/INH	Resistance detection	9
FLUOROTYPE MTB	TB detection	6
FLUROTYPE MTBDR	TB detection	4
FLUROTYPE MTBDR	Resistance detection	3
BD MAX MDR-TB	TB detection	5
BD MAX MDR-TB	Resistance detection	5
COBAS 6800/8800 MTB	TB detection	2
COBAS 6800/8800 MTB-DR	Resistance detection	2

Table 2a: Diagnostic accuracy of each index test- TB detection

Parameter	Point estimate	CI	TB+	TB -
Overall (N = 29)				
Overall sens	93.0	90.9 to 94.7	4767	9085
Overall spec	97.7	95.6 to 98.8		
Abbott RealTime MTB (N= 12)				
Pooled sens	96.1	89.9 to 98.5	1274	4162
Pooled spec	97.6	94.8 to 98.9		
FluoroType MTB (N= 6)				
Pooled sens	91.3	89.0 to 93.0	755	2502
Pooled spec	98.4	91.9 to 99.7		
FluoroType MTBDR (N= 4)				
Pooled sens	91.7	86.2 to 95.1	976	677
Pooled spec	99	95.8 to 99.8		
BD Max MTB (N= 5)				
Pooled sens	93.0	89.7 to 95.3	1098	1207
Pooled spec	95.1*	73.2 to 99.3		
Roche Cobas MTB (N= 2)**				
Sens	93.3	87.4 to 97.3	664	537
Spec	96.7	92.9 to 98.6		

CI: Confidence interval; # = number of; sens: sensitivity; spec: specificity

*A study (Paradis 2019) reported low specificity of 87%, but that study also had low specificity in Xpert and on BD Max, when a head-to-head comparison was done [72% (Bd Max) vs 73% (Xpert)]

** For Roche Cobas MTB assay, we present the results as median sensitivity and specificity. For individual studies, the sensitivity was 91.7% (92.0-97.3) and 94.9% (87.4- 94.6). the specificty was 95.6% (92.9-98.6) and 97.9% (96.2-97.8).

Table 2b: Diagnostic accuracy for TB detection by smear status

Parameter	Point estimate	CI	Point estimate	CI
Smear positive (N = 26; 2880 specimens)			**Smear negative (N = 26; 9947 specimens)**	
Overall sens	98.8%	93.7 to 99.7	86.1%	73.4 to 93.2
Overall spec	-	-	96.4%	93.6 to 98.0
Abbott RealTime MTB(N= 11)				
Pooled sens	99.7%	97.3 to 99.9	85.7%	65.5 to 95.0
Pooled spec	-	-	99.2%	96.4 to 99.1
FluoroType MTB (N= 4)				
Pooled sens	87.0%	83 to 91	84.0%	80 to 89
Pooled spec	-	-	97.0%	96 to 98
FluoroType MTBDR (N= 4)				
Pooled sens	97%	95 to 99	79.1%	62.9 to 89.4
Pooled spec	-	-	99.3%	96.2 to 99.8
BD Max MTB (N= 4)				
Pooled sens	97%	93 to 99	79.6%	68.7 to 87.4
Pooled spec	-	-	97.2%	76.6 to 99.7
Roche Cobas MTB (N= 2)*				
Sens	99%	97 to 100	80.0%	61.2 to 92.4
Spec	-	-	97.0%	96.2 to 98.6

CI: Confidence interval; # = number of; sens: sensitivity; spec: specificity

*For Roche Cobas MTB assay, data are presented as median sensitivity and specificity. For individual studies, the sensitivity in semar positive was 99% (97-99) and 99% (97-100). For smear negative the sensitivity was 73% (61.2 to 82.3), 87% (80.1 to 92.4). The specificity was 96% (96.2 to 98.6) and 98% (92.0 to 98.1).

Table 3a: Diagnostic accuracy of each index test - resistance detection with phenotypic DST as reference standard

Overall Pooled RIF (N = 18)		**Overall Pooled INH (N = 18)**	
Pooled Sens	96.7 (93.1 to 98.4)	**Pooled sens**	86.4 (82.8 to 89.3)
Pooled Spec	98.9 (97.5 to 99.5)	**Pooled spec**	99.2 (98.1 to 99.7)
Abbott (RIF) N = 9		**Abbott (INH) N = 9**	
Pooled sens	94% (91 to 96)	**Pooled sens**	89% (86 to 92)
Pooled spec	99% (99 to 100)	**Pooled spec**	98% (98 to 100)
FluoroType MTBDR (RIF) N= 3		**FluroType MTBDR (INH) N= 3**	
Sens	97 (82 – 100); 99 (94 – 100) ; 100 (93-100)	**Sens**	70 (46 – 88); 92 (84 – 97); 79 (67-88)
Spec	100 (85 – 100); 100 (96- 100) ; 95 (90-98)	**Spec**	100 (84 – 100); 100 (96-100); 99 (94-100)
BD Max MDR-TB (RIF) N= 4		**BD Max MDR-TB (INH) N= 4**	
Pooled Sens	99.1 (96.2-100)	**Pooled Sens**	90.0 (64.6 – 97.8)
Pooled Spec	98.2 (96.4 - 99.2)	**Pooled Spec**	99.8 (98.2 – 99.9)
Roche Cobas MTB RIF/INH (RIF) N= 2		**Roche Cobas MTB RIF/INH (INH) N = 2**	
Sens	91 (85-96); 100 (54-100)	**Sens**	80 (28- 99); 97 (91- 100)
Spec	96 (94-98); 100 (95-100)	**spec**	98 (96-100); 97 (91- 100)

Data are presented as point estimate with 95% CI. Sens: sensitivity; spec: specificity

Table 3b : Diagnostic accuracy of each index test - resistance detection with composite reference standard

Overall Pooled RIF (N = 9)		Overall Pooled INH (N = 8)	
Pooled sens	96.7 (92.9 to 98.5)	**Pooled sens**	86.4 (82.1 to 89.8)
Pooled spec	98.7 (97.1 to 99.4)	**Pooled spec**	99.8 (98.3 to 99.8)

Data are presented as point estimate with 95% CI. Sens: sensitivity; spec: specificity

Table 4: Sensitivity analysis

Parameter	Point estimate (95% CI)
TB detection (N = 25)	
Overall sensitivity	92.9 (90.4 to 94.8)
Overall specificity	98.1 (95.8 to 99.2)
Rifampicin resistance (N = 15)	
Overall sensitivity	97.6 (93.3 to 99.2)
Overall specificity	98.6 (97.0 to 99.4)
Isoniazid resistance (N = 15)	
Overall sensitivity	85.7 (81.3 to 89.2)
Overall specificity	99.4 (98.6 to 99.7)

CI: Confidence interval

Table 5: Summary of findings: TB detection

What is the diagnostic accuracy of cNATs for TB detection?
Patient or population: Participants with presumptive pulmonary TB
Setting: outpatient and inpatient
Reference standard: culture
Pooled sensitivity: 0.93 (95% CI: 0.91 to 0.95) **Pooled specificity:** 0.98 (95% CI: 0.96 to 0.99)

Test result	Number of results per 1,000 patients tested (95% CrI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 2.5%	Prevalence 10%	Prevalence 30%		
True positives	23 (23 to 24)	93 (91 to 95)	279 (273 to 284)	4767 (29)	⊕⊕⊕◯ Moderate a, b
False negatives	2 (1 to 2)	7 (5 to 9)	21 (16 to 27)		
True negatives	953 (932 to 963)	879 (860 to 889)	684 (669 to 692)	9085 (29)	⊕⊕⊕⊕ High
False positives	22 (12 to 43)	21 (11 to 40)	16 (8 to 31)		

CrI: Credible interval

Explanations

a. Of the total 29 studies, 16 (55%) had high or unclear risk of bias as they either did prior testing before including specimens in the study or used convenience sampling or the method of participant selection was not reported. We downgraded one level for risk of bias.
b. Median TB prevalence in these studies was 31% and the number of specimens for TB positive and TB negative are large, so we decided to not downgrade for indirectness.

Table 6: Summary of findings: cNATs for rifampicin resistance

What is the diagnostic accuracy of cNATs for RIF resistance detection?
Patient or population: Participants with presumptive MDR-TB
Setting: outpatient and inpatient
Reference standard: **culture**
Pooled sensitivity: 0.97 (95% CI: 0.93 to 0.98) | **Pooled specificity** : 0.99 (95% CI: 0.97 to 0.99)

Test result	Number of results per 1,000 patients tested (95% CI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 2% Typically seen in	**Prevalence 10%** Typically seen in	**Prevalence 15%** Typically seen in		
True positives	**19** (19 to 20)	**97** (93 to 98)	**145** (140 to 148)	702 (18)	⊕⊕⊕◯ Moderate [a,b]
False negatives	**1** (0 to 1)	**3** (2 to 7)	**5** (2 to 10)		
True negatives	**969** (956 to 975)	**890** (878 to 896)	**841** (829 to 846)	2172 (18)	⊕⊕⊕⊕ High
False positives	**11** (5 to 24)	**10** (4 to 22)	**9** (4 to 21)		

CrI: Credible interval

a. There were 8 (44%) out of 18 studies that had high or unclear risk of bias as the participant selection was not reported or there was prior testing done for the specimens included in the study. We downgraded one level for risk of bias.
b. The median prevalence of rifampicin resistance in these studies was 15%, which is representative of drug resistance in most countries for pulmonary TB. We did not downgrade for indirectness

Table 7: Summary of findings: cNATs assay for isoniazid resistance

What is the diagnostic accuracy of cNATs assay for INH resistance detection?
Patient or population: Participants with presumptive MDR-TB
Setting: outpatient and inpatient
Reference standard: culture
Pooled sensitivity: 0.86 (95% CI: 0.83 to 0.89) | **Pooled specificity** : 0.99 (95% CI: 0.98 to 1.00)

Test result	Number of results per 1,000 patients tested (95% CrI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 2% Typically seen in	**Prevalence 10%** Typically seen in	**Prevalence 15%** Typically seen in		
True positives	**17** (17 to 18)	**86** (83 to 89)	**130** (124 to 134)	854 (18)	⊕⊕⊕◯ **Moderate** [a,b,c]
False negatives	**3** (2 to 3)	**14** (11 to 17)	**20** (16 to 26)		
True negatives	**972** (961 to 977)	**893** (883 to 897)	**843** (834 to 847)	1904 (18)	⊕⊕⊕⊕ **High** [c]
False positives	**8** (3 to 19)	**7** (3 to 17)	**7** (3 to 16)		

CrI: Credible interval

a. There were 8 (44%) out of 18 studies that had high or unclear risk of bias as the participant selection was not reported or there was prior testing done for the specimens included in the study. We downgraded one level for risk of bias.
b. Sensitivity for INH resistance ranges from 58% to 100%. There was one study with low sensitivity, however, overlapping confidence intervals were seen. We did not downgrade for inconsistency.
c. The median prevalence in these studies was 19.7%. With high number of specimens being evaluated in these studies, we did not downgrade for indirectness.

Table 8: Summary of findings table: cNATs for RIF resistance: CRS

What is the diagnostic accuracy of cNATs for RIF resistance detection?
Patient or population: Participants with presumptive pulmonary TB
Setting: outpatient and inpatient
Reference standard: CRS: phenotypic DST (solid or liquid) and gene sequencing, where at least one test is resistant
Pooled sensitivity : 0.97 (95% CI: 0.93 to 0.98) | **Pooled specificity** : 0.99 (95% CI: 0.97 to 0.99)

Test result	Number of results per 1,000 patients tested (95% CrI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 2% Typically seen in	Prevalence 10% Typically seen in	Prevalence 15% Typically seen in		
True positives	**19** (19 to 20)	**97** (93 to 99)	**145** (139 to 148)	565 (9)	⊕⊕⊕◯ **Moderate** [a,b]
False negatives	**1** (0 to 1)	**3** (1 to 7)	**5** (2 to 11)		
True negatives	**967** (952 to 974)	**888** (874 to 895)	**839** (825 to 845)	1120 (9)	⊕⊕⊕⊕ **High**
False positives	**13** (6 to 28)	**12** (5 to 26)	**11** (5 to 25)		

CrI: Credible interval

Explanations

a. There were 6 (66%) out of 9 studies that had high or unclear risk of bias as the participant selection was not reported or there was prior testing done for the specimens included in the study. We downgraded one level for risk of bias.
b. The median prevalence of rifampicin resistance in these studies was 36%. We did not downgrade for indirectness

Table 9: Summary of findings table: cNATs for INH resistance: CRS

What is the diagnostic accuracy of cNATs for INH resistance detection?
Patient or population: Participants with presumptive pulmonary TB
Setting: outpatient and inpatient
Reference standard: CRS: phenotypic DST (solid or liquid) and gene sequencing, where at least one test is resistant
Pooled sensitivity : 0.86 (95% CI: 0.82 to 0.90) | **Pooled specificity** : 1.00 (95% CI: 0.98 to 1.00)

Test result	Number of results per 1,000 patients tested (95% CrI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 2% Typically seen in	Prevalence 10% Typically seen in	Prevalence 15% Typically seen in		
True positives	**17** (16 to 18)	**86** (82 to 90)	**130** (123 to 135)	656 (9)	⊕⊕⊕◯ **Moderate** [a,b]
False negatives	**3** (2 to 4)	**14** (10 to 18)	**20** (15 to 27)		
True negatives	**978** (963 to 978)	**898** (885 to 898)	**848** (836 to 848)	787 (9)	⊕⊕⊕⊕ **High**
False positives	**2** (2 to 17)	**2** (2 to 15)	**2** (2 to 14)		

CrI: Credible interval

Explanations

a. There were 6 (66%) out of 9 studies that had high or unclear risk of bias as the participant selection was not reported or there was prior testing done for the specimens included in the study. We downgraded one level for risk of bias.
b. The median prevalence of rifampicin resistance in these studies was 45%, which is representative of drug resistance in most countries for pulmonary TB. We did not downgrade for indirectness

Figures

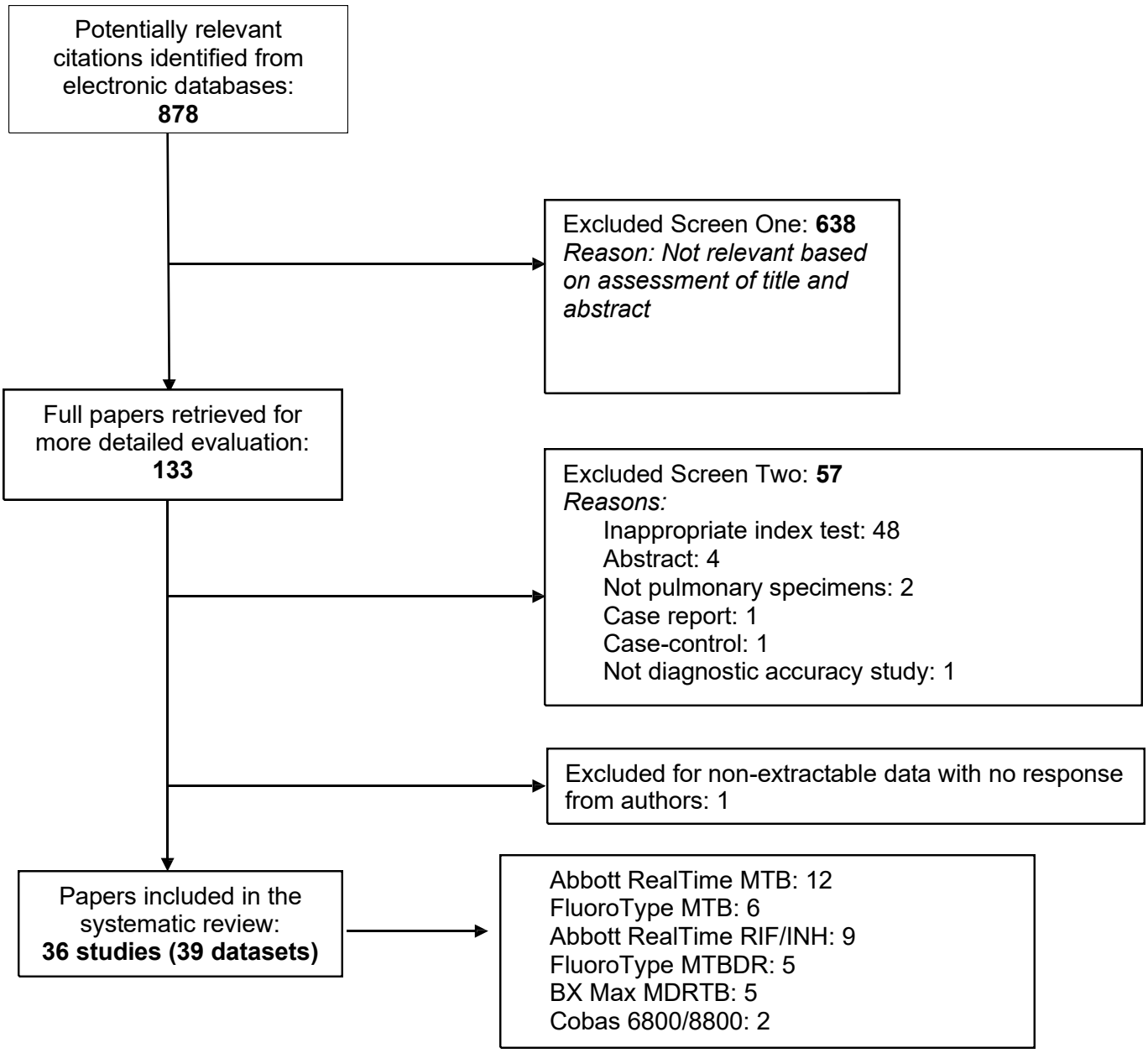


Figure 1. PRISMA diagram of studies included in the review

Appendix A: Quallity assessment

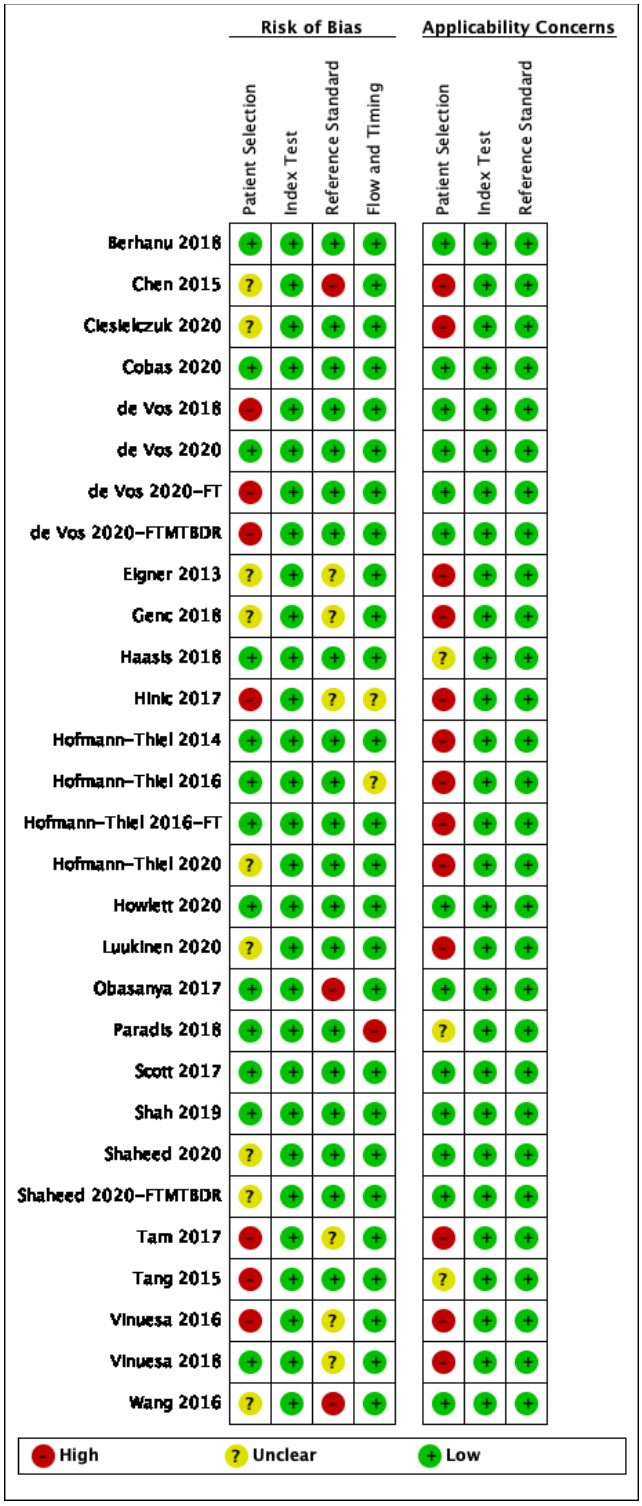


Figure A. Risk of Bias and Applicability Concerns summary for QUADAS-2 domains in each study evaluating assays for TB detection

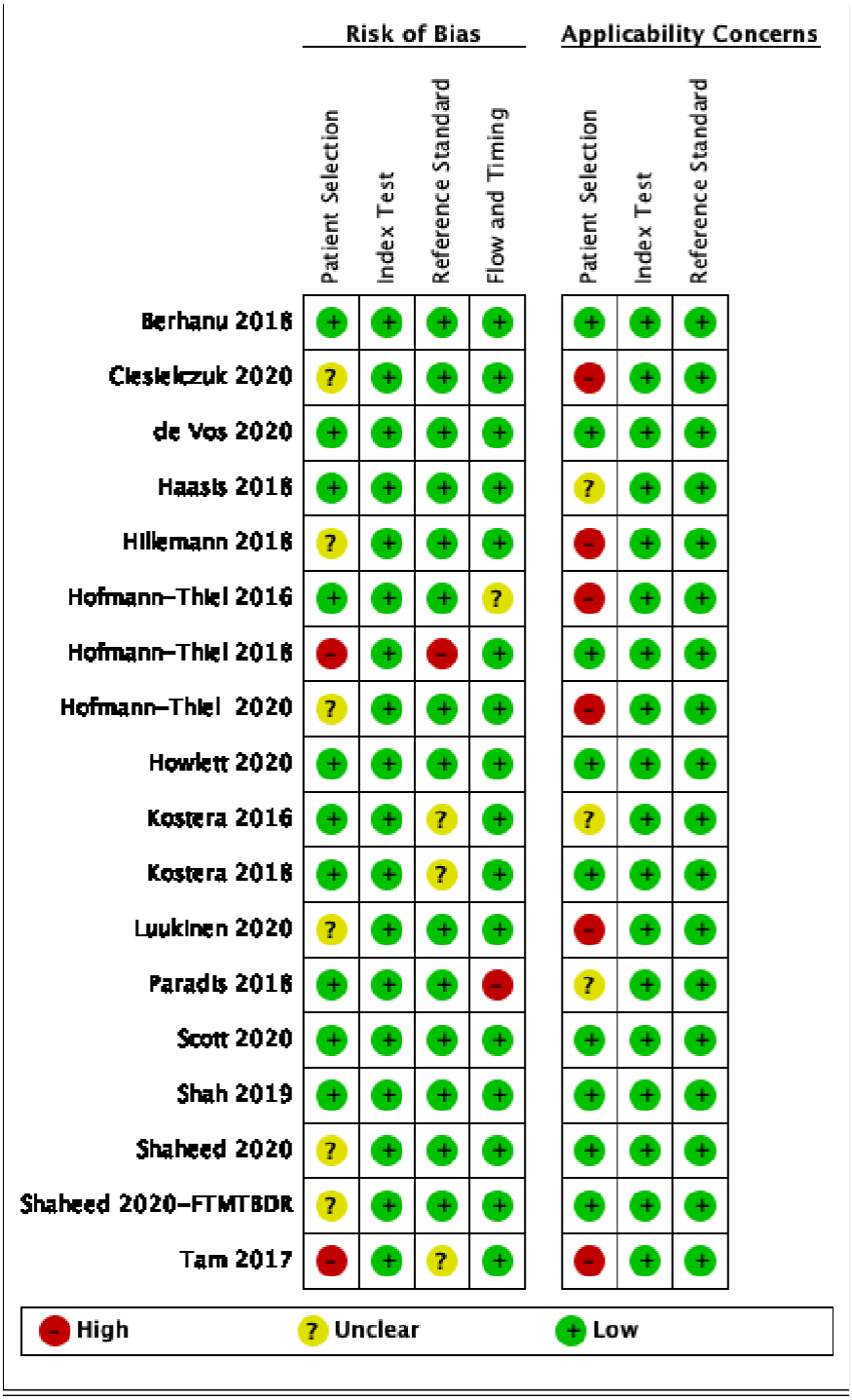


Figure B: Risk of Bias and Applicability Concerns summary for QUADAS-2 domains in each study evaluating assays for resistance detection

Appendix B. QUADAS-2 Protocol

Domain 1 Patient Selection:

Risk of Bias: Could the selection of patients have introduced bias?

- Signaling question 1: Was a consecutive or random sample of patients or specimens enrolled?
    - We scored 'yes' if the study enrolled a consecutive or random sample of eligible patients; 'no' if the study selected patients by convenience, and 'unclear' if the study did not report the manner of patient selection or this cannot be discerned.

- Signaling question 2: Was a case-control design avoided?
    - We scored 'yes' if the study enrolled only patients presumed of drug-resistant TB, including patients with confirmed TB. We scored 'no' if the study enrolled patients for whom resistance status was already known, and 'unclear' if the study did not report the design or this cannot be discerned.
- Signaling question 3: Did the study avoid inappropriate exclusions?
    - We scored 'yes' if no inappropriate exclusions were noted. We scored 'no' if studies note specific exclusions. Inappropriate exclusions could potentially occur if patients were excluded based on prior knowledge or testing about them or if the technician does not record performed test results but this was not anticipated for research studies in this review.

Applicability: Are there concerns that the included patients and setting do not match the review question?

We were interested in how the index tests (centralized molecular DST assays) performed in patients presumed of having TB who are evaluated. We judged 'low concern' when the specimens included in the study were from the patients with presumptive pulmonary TB and was conducted in high TB and/or high MDR-TB burden country as per the WHO list. We judged 'high concern' if the specimens were collected from patients in a low TB and/or MDR-TB burden country. We will judge 'unclear concern' if the study included specimens from both high and low TB/MDR-TB burden settings or we could not tell.

Domain 2: Index Test

Risk of Bias: Could the conduct or interpretation of the index test have introduced bias?

- Signaling question 1: Were the index test results interpreted without knowledge of the results of the reference standard?
    - We scored 'yes' for all studies because all the centralized molecular DST assay results are automatically generated and the user is provided with printable test results. Thus, there was no room for subjective interpretation of test results.

- Signaling question 2: If a threshold was used, was it prespecified?

- As the threshold is prespecified in all centralized molecular DST assay in this review, we answered this question "yes" for all studies.

Applicability: Are there concerns that the index test, its conduct, or its interpretation differ from the review question? Variations in test technology, execution, or interpretation may affect estimates of the diagnostic accuracy of a test.

We judged 'low concern' if the test was done as per recommendation of the manufacturer for PTB specimens. We judged 'high concern' it was stated and/or if additional steps were used for sample preparation and 'unclear concern' if we could not tell.

Domain 3: Reference Standard

Risk of Bias: Could the reference standard, its conduct, or its interpretation have introduced bias?

- Signaling question 1: Is the reference standard likely to correctly classify the target condition?
  - For detection of TB, culture is generally considered the best reference standard. We scored 'yes' if the studies used MGIT 960 as the reference standard (higher quality reference standard). We scored 'no' if the studies used only solid media-based culture (lower quality reference standard) as all these index tests are for centralized settings, we expect the laboratory settings to have liquid culture for detecting TB. LJ culture has lower diagnostic accuracy than liquid culture and would over or under-estimate the diagnostic accuracy of the index test. We scored 'unclear' if we could not tell.
  - For detection of rifampicin resistance, culture-based drug susceptibility testing (DST, also called conventional phenotypic method) is considered to be the best reference standard. As we extracted data for studies that used culture-based DST, we will score " yes" for all studies.

- Signaling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?
  - We scored 'yes' if the reference test provided was culture e.g. MGIT 960 DST where an automated result is generated (except for LJ with confirmation of MTB by a NAAT-based test), if blinding was explicitly stated, or if it was clear that the reference standard was performed at a separate laboratory and/or performed by different people. We will score 'no' if the study stated that the reference standard was interpreted with knowledge of the index test result. We scored 'unclear' if this was not stated or answered inadequately.

- Signaling question 3: (Rifampicin resistance) Were the reference standard results interpreted without knowledge of the results of the index test?

- We added a signaling question for rifampicin resistance detection. We scored "yes" if the reference test provided an automated result (for example, MGIT 960), blinding was explicitly stated, or it was clear that the reference standard was performed at a separate laboratory or performed by different people, or both. We scored "no" if the study stated that the reference standard result was interpreted with knowledge of the index test result. We scored "unclear" if we could not tell.

Applicability: Are there concerns that the target condition as defined by the reference standard does not match the question?
We judged applicability to be of 'low concern' for all studies.

Domain 4: Flow and Timing
Risk of Bias: Could the patient flow have introduced bias?

- Signaling question 1: Was there an appropriate interval between the index test and reference standard?
  - We scored ''yes' if the tests were paired or separated by less than 48 hours after treatment initiation. We scored 'no' if the reference and index tests were not performed on paired specimens or were separated by more than a week. We scored 'unclear' if this was not stated in the paper or answered inadequately. In the majority of included studies, we expected specimens for index tests and culture to be obtained at the same time (i.e. to be performed on paired specimens for the majority of studies), when patients are presumed of having TB or MDR-TB.

- Signaling question 2: Did all patients receive the same reference standard?
  - For the diagnosis of TB, we scored this question "yes" if all participants in the study or a subset of participants in the study (for whom we will extract data) received the acceptable reference standard (solid culture, liquid culture, or both), which we specified as a criterion for inclusion in the review. However, we acknowledge that it is possible that some specimens could undergo solid culture and others liquid culture as the reference standard. This variation was recorded.
  - For rifampicin resistance detection, we scored "yes" if all participants received the same reference standard (either culture-based DST or MTBDR*plus*), "no" if not all participants received the same reference standard, and "unclear" if we could not tell.

- Signaling question 3: Were all patients included in the analysis?
- The answer to this question was determined by comparing the number of patients enrolled with the number of patients included in the two-by-two tables. We noted if authors record the number of indeterminate results. We scored 'yes' if the number of participants enrolled was clearly stated and corresponded to the number presented in the analysis or if exclusions were adequately described. We scored 'no' if there were participants missing or excluded from the

analysis and there was no explanation given; and 'unclear ' if not enough information was given to assess whether participants were excluded from the analysis

Appendix C: Data by different technologies

TB detection

1. Abbott RealTime MTB assay

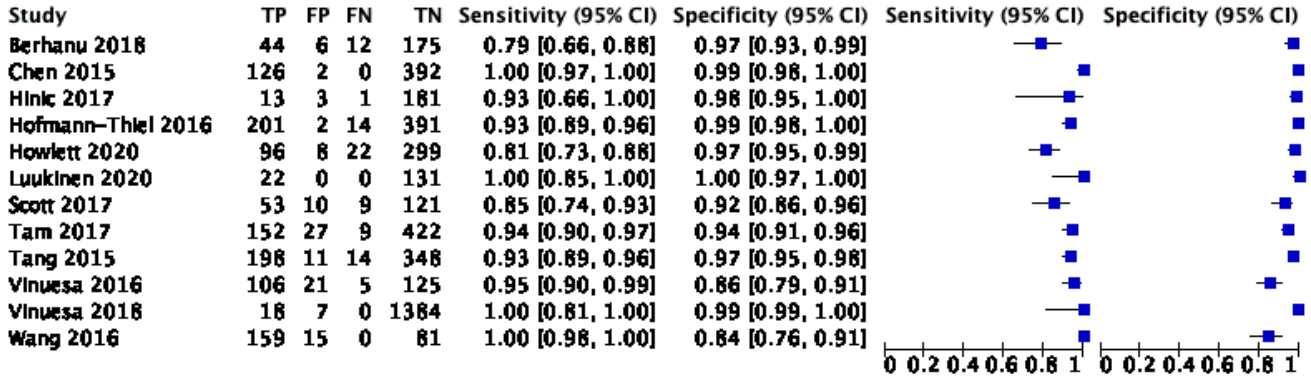


Figure A: Forest plot of TB detection by Abbott RealTime MTB assay

2. FluoroType MTB assay

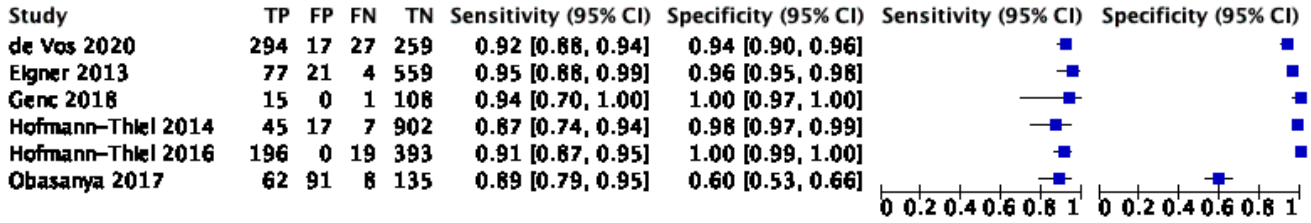


Figure B: Forest plot of TB detection by FluoroType MTB assay

3. FluoroType MTBDR assay

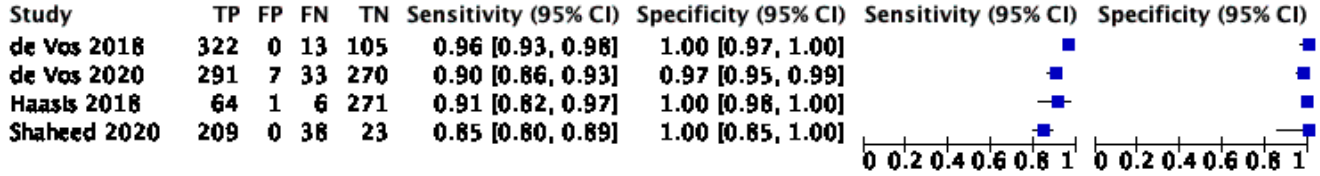


Figure C: Forest plot of TB detection by FluoroType MTBDR assay

4. BD Max MDR-TB

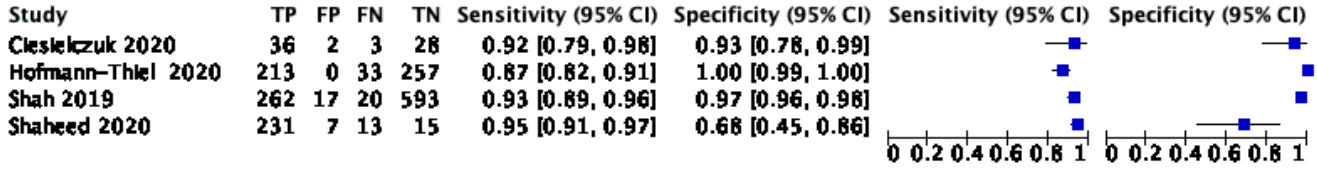


Figure D: Forest plot of TB detection by BD Max MDR-TB assay

5. Cobas 6800/8800 MTB assay

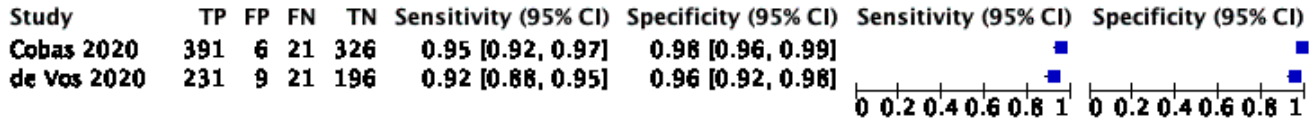


Figure E: Forest plot of TB detection by Cobas MTB assay

Resistance detection

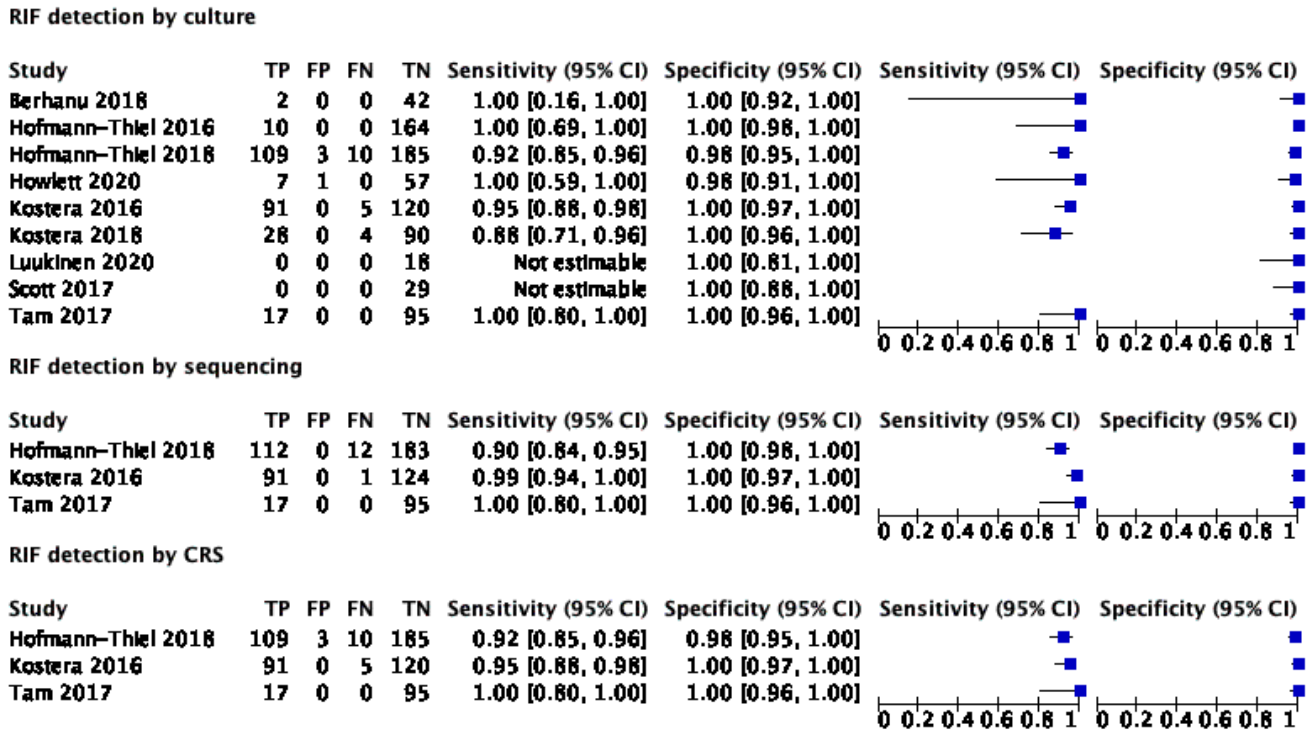


Figure: RIF resistance detection by Abbott RealTIme RIF/INH resistance

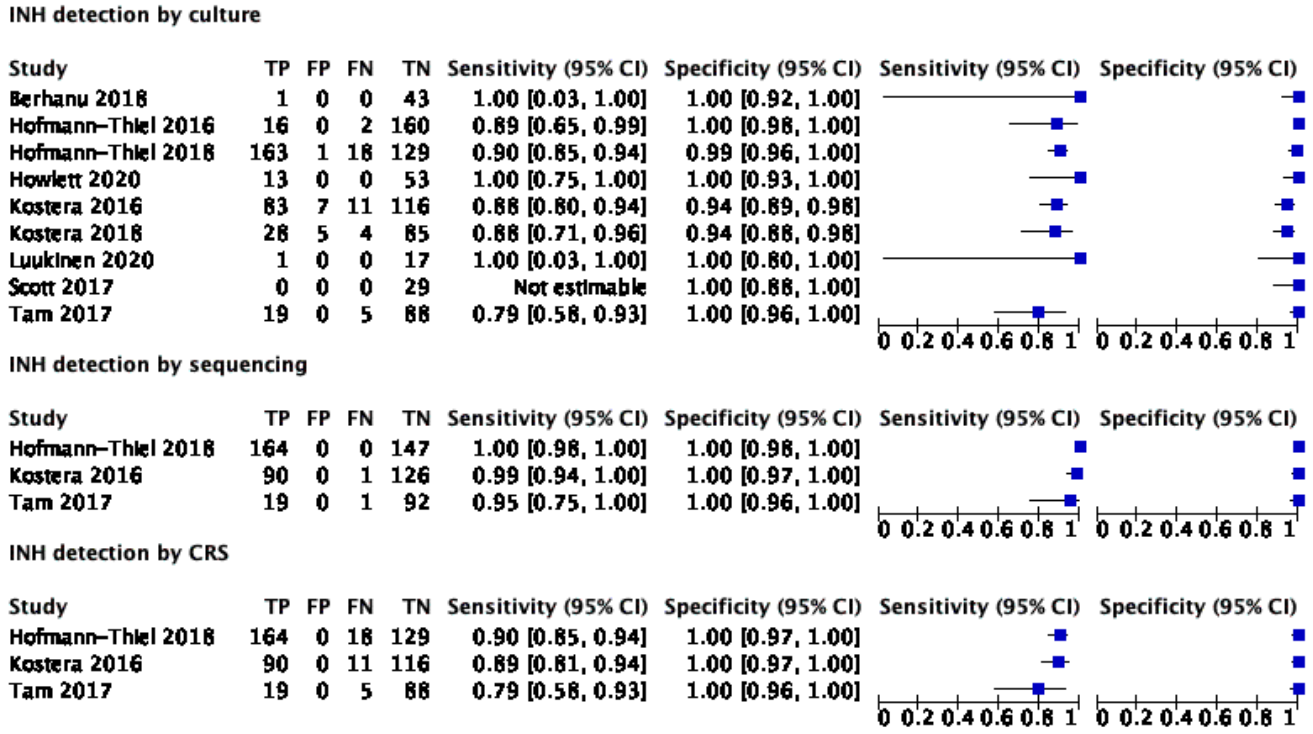


Figure: INH resistance detection by Abbott RealTIme RIF/INH resistance

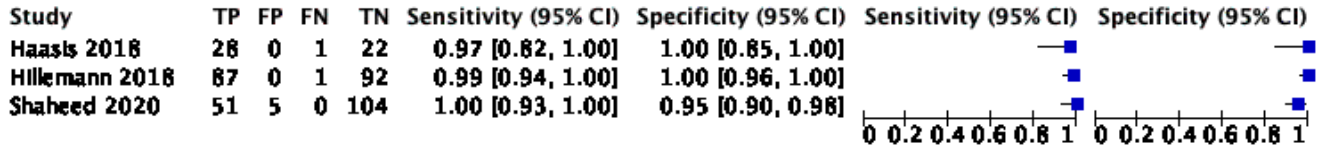


Figure: RIF resistance detection by FluroType MTBDR resistance

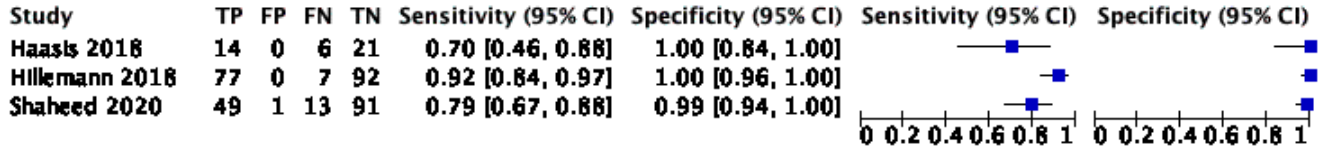


Figure: INH resistance detection by FluroType MTBDR resistance

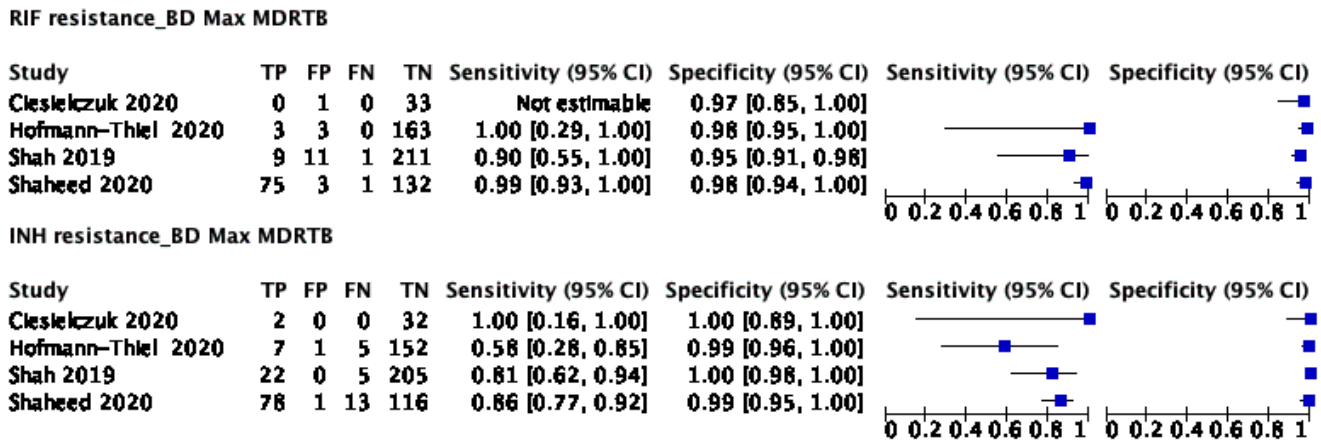


Figure: RIF and INH resistance detection by BD Max MDR-TB assay

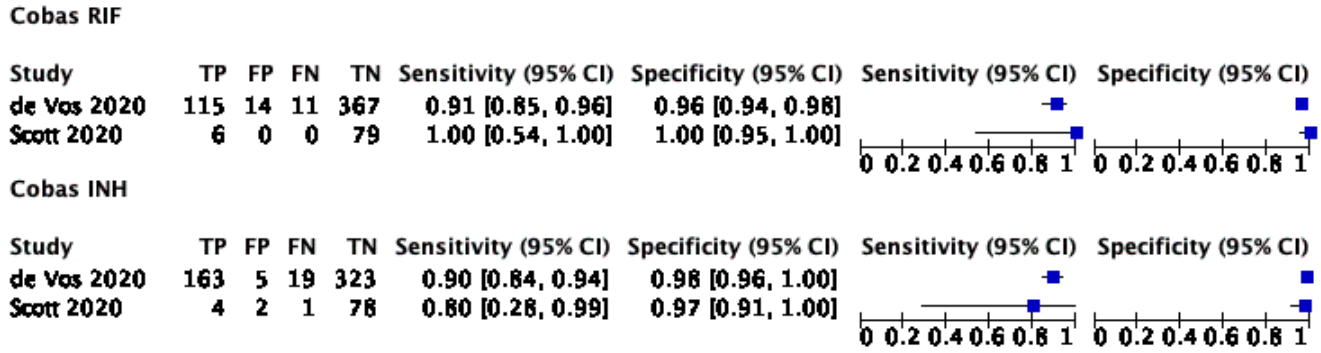


Figure: RIF and INH resistance detection by Cobas MTB-DR assay

Appendix D

Composite reference standard

Three studies (Hofmann Thiel, 2018; Kostera 2016 and Tam 2017) evaluated sequencing and provided information on the mutations to assess results using CRS for RIF resistance. In the study by Hofmann Thiel 2018, three specimens which were susceptible by phenotypic DST but resistant by Abbott, had a L511P mutation as confirmed by sequencing. As there is limited confidence that this mutation is associated with resistance (Miotto 2017), we interpreted the composite reference standard as sensitive and thus classified these results as false positive. Ten specimens which were resistant by phenotypic DST but susceptible by index test were confirmed to be resistant by sequencing as well. As we had high confidence that mutations in these 10 specimens (6 specimens with H526R and 4 with L533P mutations) were associated with resistance, we classified them as false-negative with CRS. Additionally, two specimens had a silent mutation K527K in the *rpoB* gene, which were susceptible on culture media. Hence, the sensitivity for RIF resistance with CRS was identical to phenotypic DST (92%) but decrease with sequencing to 90%. Similarly, for specificity, with phenotypic and CRS as reference standard, it was identical (98%), but increased with sequencing to 100%.

For Kostera 2016, of the five discordant specimens, which were resistant by the phenotypic DST but susceptible by index test, four were confirmed to be wild type by sequencing. However, with CRS, these four specimens were considered resistant, as the mutations might have been out of the target region of Sanger sequencing. Hence, the sensitivity for RIF resistance with CRS was 95%, identical to phenotypic DST but it increased to 99% with sequencing. Specificity was similar with all three reference standard (100%).

In the Hillemann 2018 study, FluoroType MTBDR missed mutations in both *katG* and *inhA* promoter region. The assay missed S315T1 mutation in *katG* gene, C-17T and T- 8A mutations in the inhA promoter region.

References

Abbott 2019a
Abbott. Abbott RealTime MTB. https://www.molecular.abbott/int/en/products/infectious-disease/realtime-mtb 2019.

Abbott 2019b
Abbott. Abbott RealTime RIF/INH. https://www.molecular.abbott/int/en/products/infectious-disease/realtime-mtb-rif-inh-resistance 2019.

Almeida Da Silva 2011
Almeida Da Silva PE and Palomino JC. Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs. J Antimicrob Chemother 2011;66(7):1417-1430.

Bruker-Hain Diagnostics 2019a
Bruker-Hain Diagnostics. FluoroType MTB. https://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/fluorotype-mtb.html 2019.

Bruker-Hain Diagnostics 2019b
Bruker-Hain Diagnostics. FluoroType MDRTB. https://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/fluorotype-mtbdr.html 2019.

Burdz 2003
Burdz TV, Wolfe J, Kabani A. Evaluation of sputum decontamination methods for Mycobacterium tuberculosis using viable colony counts and flow cytometry. Diagn Microbiol Infect Dis , 2003; 47 :503-9.

Chihota 2010
Chihota VN, Grant AD, Fielding K, Ndibongo B, van Zyl A, et al. Liquid vs. solid culture for tuberculosis: performance and cost in a resource-constrained setting. Int J Tuberc Lung Dis , 2010; 14: 1024-1031.

Dowdy 2008
Dowdy DW, Lourenco MC, Cavalcante SC, Saraceni V, King B, et al. Impact and cost-effectiveness of culture for diagnosis of tuberculosis in HIV-infected Brazilian adults. PLoS One, 2008; 3: e4057.

Chu 2006

Chu H and Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol 2006;59(12):1331-1332.

Cochrane 2008
The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. 2008.

Dicks 2019
Dicks KV and Stout JE. Molecular Diagnostics for Mycobacterium tuberculosis Infection [Annual Review of Medicine]. 2019;70(1):77-90.

Friedrich 2011
Friedrich SO, Venter A, Kayigire XA, Dawson R, Donald PR, Diacon AH. Suitability of Xpert MTB/RIF and genotype MTBDRplus for patient selection for a tuberculosis clinical trial. *Journal of Clinical Microbiology* 2011;**49**(8):2827–31.

GLI 2014
Global Laboratory Initiative. Mycobacteriology Laboratory Manual, 2014. https://www.who.int/tb/laboratory/mycobacteriology-laboratory-manual.pdf (accessed 10 March 2019).

Horne 2019
Horne DJ, Kohli M, Zifodya JS, Schiller I, Dendukuri N, Tollefson D et al. Xpert MTB/RIF and Xpert MTB/RIF Ultra for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database of Systematic Reviews 2019, Issue 6. Art No. CD009593.

Miotto 2017
Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. Eur Resp J 2017;50(6):1701354.

Miotto 2018
Miotto P, Cabibbe AM, Borroni E, Degano M, Cirillo DM. Role of Disputed Mutations in the rpoB Gene in Interpretation of Automated Liquid MGIT Culture Results for Rifampin Susceptibility Testing of Mycobacterium tuberculosis. J Clin Microbiol. 2018; 56(5): e01599-17.

Moher 2009
Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097.

Nathavitharana 2017
Nathavitharana RR, Cudahy PG, Schumacher SG, Steingart KR, Pai M, Denkinger CM. Accuracy of line probe assays for the diagnosis of pulmonary and multidrug-resistant tuberculosis: a systematic review and meta-analysis. The European respiratory journal 2017;49(1):1601075. DOI: 10.1183/13993003.01075-2016

Reitsma 2005
Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982-990.

Schünemann 2020
Schünemann HJ, Mustafa R, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy. Journal of Clinical Epidemiology 2020. [DOI: 10.1016/j.jclinepi.2019.12.020]

Schünemann 2020
Schünemann HJ, Mustafa R, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. Inconsistency, Imprecision, publication bias and other domains for rating the certainty of evidence for test accuracy and presenting it in evidence profiles and summary of findings tables. Journal of Clinical Epidemiology 2020;pii: S0895-4356(19)30674-2. [DOI: 10.1016/j.jclinepi.2019.12.021]

Smith 2012
Smith SE, Kurbatova EV, Cavanaugh JS, Cegielski JP. Global isoniazid resistance patterns in rifampin-resistant and rifampin-susceptible tuberculosis.. Int J Tuberc Lung Dis 2012;16(2):203-205.

Somoskovi 2001
Somoskovi A, Parsons LM, Salfinger M.. The molecular basis of resistance to isoniazid, rifampin, and pyrazinamide in Mycobacterium tuberculosis.. Respiratory research 2001;2(3):164-168.

Subbaraman 2016
Subbaraman R, Nathavitharana RR, Satyanarayana S, Pai M, Thomas BE, et al.. The Tuberculosis Cascade of Care in India's Public Sector: A Systematic Review and Meta-analysis. PLOS Medicine 2016;13(10):e1002149.

Theron 2016
Theron G, Venter R, Calligaro G, Smith L, Limberis J, Meldau R, et al. Xpert MTB/RIF Results

in Patients With Previous Tuberculosis: Can We Distinguish True From False Positive Results?. Clinical Infectious Diseases 2016;62 (8):995–1001.

Theron 2016
Theron G, Venter R, Calligaro G, Smith L, Limberis J, Meldau R, et al. Xpert MTB/RIF Results in Patients With Previous Tuberculosis: Can We Distinguish True From False Positive Results?. Clinical Infectious Diseases 2016;62 (8):995–1001.

Theron 2018
Theron G, Venter R, Smith L, Esmail A, Randall P, Sood V, et al. False positive Xpert MTB/RIF results in re-tested patients with previous tuberculosis: frequency, profile, and prospective clinical outcomes. Journal of Clinical Microbiology 2018;56(3):pii: e01696-17.

UNITAID 2017
UNITAID. TUBERCULOSIS: Diagnostics Technology Landscape. UNITAID 2017.

Wang 2019
Wang MG, Xue M, Wu SQ, Zhang MM, Wang Y, Liu Q et al. Abbott RealTime MTB and MTB RIF/INH assays for the diagnosis of tuberculosis and rifampicin/isoniazid resistance. Infection, Genetics and Evolution 2019; 71: 54-59.

Whiting 2011
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(8):529-536.

WHO 2014
World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. Geneva, World Health Organization 2014.

WHO 2018
World Health Organization. Global Tuberculosis Report 2017. Geneva, World Health Organization 2017.

WHO 2018
World Health Organization. Treatment guidelines for isoniazid-resistant tuberculosis: Supplement to the WHO treatment guidelines for drug-resistant tuberculosis 2018.

Web Annex D.9. Systematic Literature Review of Economic Evidence for Nucleic acid amplification tests (NAATs) to detect TB and DR-TB in adults and children

Produced in Preparation for the WHO guideline development group meeting Dec. 7-18, 2020

Final Report: PLEASE DO NOT DISTRIBUTE
December 2020

Alice Zwerling, Sazia Naz & Brianna Empringham
School of Epidemiology & Public Health
Faculty of Medicine,
University of Ottawa, Ottawa, Canada
azwerlin@uottawa.ca

BACKGROUND

In preparation for the upcoming guideline development group meeting "Nucleic acid amplification tests to detect TB and DR-TB" scheduled for 7-11 December, 2020, there is a need to summarize the current economic evidence on the following three classes of NAATs:

1) Centralized assays that present end-to-end (E2E) solutions for detection of pulmonary TB and resistance to rifampicin and isoniazid; including the following commercially available platforms: Abbott RealTime MTB, Abbott RealTime RIF/INH, FluoroType MTB, FluoroType MTDBR, BD Max MDR-TB assay, Roche Cobas MTB and MTB-RIF/INH, and Bioneer AccuPower® TB & MDR Real-Time PCR Kit.

2) Cartridge-based nucleic acid amplification tests (CBNAT) for detection of isoniazid and second-line drug resistance, XDR-TB; including Xpert MTB/XDR and MeltPro by Zeesan.

3) Hybridization-based technology for pyrazinamide resistance detection (PZA LPA): Genoscholar PZA LPA, Nipro

OBJECTIVE

To perform a systematic review of the published literature on economic evaluations on the three above-mentioned NAATs/classes of NAATs to detect TB and DR-TB. To summarize current economic evidence and further understand the costs, cost-effectiveness of these NAATs for TB diagnosis.

For above-mentioned NAATs/classes of NAATs

1. How large are the resource requirements (costs)?
2. What is the certainty of the evidence of resource requirements (costs)?
3. Does the cost-effectiveness of the intervention favor the intervention or the comparison (Xpert MTB/RIF/Culture/Sequencing) as applicable?

Secondary objective will be to describe important variability in costs and cost-effectiveness within each class of technology.

METHODS

Search strategy & Data Sources

We performed a search of four online databases: EMBASE, Medline, Web of Science and Scopus for new studies published from January 1, 2010 through September 17th, 2020. We reviewed citations of all eligible articles, guidelines and reviews for additional studies. We also contacted experts and test manufacturers to identify any additional unpublished studies.

The search strategy used was modified to meet the criteria of each database but generally included the following terms and structure: specific search terms associated with diagnostic tests under investigation AND (tuberculosis OR TB OR mycobacterium) *AND* ("cost-benefit*" OR cost* OR economic* OR "cost effectiveness*" OR "cost-utility" OR "disability adjusted life year*" OR DALY OR "quality-adjusted life year*" OR QALY OR "cost benefit analysis" OR "cost effectiveness analysis" OR "quality of life" OR "utility").

Due to limited direct evidence, several subsequent searches were performed in the same 4 databases as of late November to identify indirect evidence using these platforms in non-TB disease areas. Search strategies were run with the specific platform names but without the following search term for TB: (tuberculosis OR TB OR mycobacterium). Additional searches were also run with additional search terms for Genotype MTBDRsl and MTBDRplus to look for indirect evidence from line probe assays to inform the evidence around PZA LPA.

Study selection & data extraction

Studies were included if they touched on any economic evidence directly related to test or test implementation costs. Studies using both primary and secondary cost data sources were included and studies from any patient population or country were considered.

The study selection followed PRISMA guidelines(Liberati et al., 2009; Moher et al., 2009). Potentially relevant studies were identified through electronic searches of the online databases as described above, and duplicates were removed. An initial abstract review of each study was completed independently by two reviewers (SN & BE); articles were excluded if they did not evaluate one of our diagnostic tests, or if they were reviews, letters or opinion pieces (i.e. no original data), conference abstracts were also excluded. Full text review was then completed on remaining articles, and articles that met predetermined inclusion criteria were retained for the review.

Full texts of included studies including published supplemental material, were independently reviewed by two reviewers (SN, BE), with all disagreements resolved by discussion with a third reviewer (AZ).

The study design data elements extracted from each study included: the primary research question, country and setting, year of study, patient population, clinical setting, diagnostic scenarios, comparison diagnostic scenarios and reference scenarios, economic analysis perspective, analytic time horizon, type of economic evaluation, source of costing, primary outcome measure, type of model, types of sensitivity and uncertainty analyses performed and willingness-to-pay threshold.

Cost components and unit test costs were extracted, along with key costing input parameters. Costs are presented in USD (United States Dollars) unless otherwise noted.

RESULTS

End-to-end solutions (E2E) for detection of pulmonary TB (TB), RIF and INH resistance – PICO 1

Several commercially available tests were included as eligible tests in the E2E category, however no published studies were identified assessing the costs or cost-effectiveness of any E2E solutions. One unpublished study from FIND looking at the BD MAX and Hain assays was identified and the data is described below.

Unpublished data from FIND was provided through direct communication, this costing-only study used time and motion studies combined with a bottom-up, ingredients-based approach to estimate the unit test cost for the BD Max and Hain test respectively. Time and motion studies were conducted at a reference level laboratory in South Africa. Time-and-Motion studies typically involve direct observation

of staff conducting various activities to capture discrete time estimates that can be attributed to each activity and used to construct personnel cost per activity. Those staff costs are then used in the bottom-up ingredients approach where component costs (staff, consumables, equipment, etc.) are added up to create a total unit cost. Several important simplifying assumptions were made that may limit generalizability of results, including assuming 50% of lab operations were dedicated to TB, a minimum daily throughput of 24 samples/day or the equivalent of one BD MAX run (24 test/run), equipment costs were fixed at $100,000 for both machines and a 5% annual maintenance cost was assumed and the standard 3% discount rate and 10 years expected useful life years. Key parameter values and low and high values used to calculate overhead, building, staff unit test costs are included in table 1, all costs from this analysis are presented in 2019 USD.

Table 1. Key parameter values and ranges for unit test cost calculations

Key parameters	PE	Low	High
Approximate Annual Operational days	250	300	200
Hours per operating day	8	10	6
Laboratory Technician Salary (Annual)	$5,000	$2,500	$7,500
Annaul building cost*	$2,265	$520	$6,955
Laboratory overhead (utility + asset + staff)**	$65,500	$32,750	$98,250

Data courtesy of FIND/H Sohn, unpublished. PE: Primary estimate

Unit test costs were estimated at $18.52 per test (Range: $13.79-$40.70) for BD MAX and $15.37 per test ($9.61- $37.40) for Hain. Hain is approximately $4 cheaper per test kit but with higher operational costs. Differences across the two tests in unit test cost are driven primarily by staff costs which are higher for Hain ($0.77 vs $1.66) and a higher test kit price for BD MAX ($13 vs $9). Unit test component costs are provided in tables 2A & B.

Table 2A. Components of unit test cost for BD MAX from unpublished FIND data

Type of test:	BD		
Resource Type	**PE**	**Low**	**High**
Overhead	$1.46	$0.48	$3.71
Building	$0.39	$0.06	$2.00
Staff	$0.77	$0.32	$1.45
Equipment	$2.89	$1.93	$18.53
Test Kits	$13.00	$11.00	$15.00
Total	$18.52	$13.79	$40.70

Data courtesy of FIND/H Sohn, unpublished. PE: Primary estimate

Table 2B. Components of unit test cost for Hain from unpublished FIND data

Type of test:	Hain		
Resource Type	**PE**	**Low**	**High**
Overhead	$1.95	$0.64	$5.00
Building	$0.33	$0.05	$1.69
Staff	$1.66	$0.69	$3.11
Equipment	$2.43	$1.62	$15.60
Test Kits	$9.00	$6.60	$12.00
Total	$15.37	$9.61	$37.40

Data courtesy of FIND/H Sohn, unpublished. PE: Primary estimate

The cost of the test kit was the largest single cost contributor across both tests (~$4 less for Hain kit) however under conditions with higher annual overhead costs and fewer annual operational days and operational hours/day ('HIGH'er GDP scenario) equipment costs became the largest single cost component under these conditions, as seen in Figures 2A & B. Equipment costs were fixed in this analysis at $100,000 and was a strong driver of cost variation depending on utilization, which will vary across different laboratory networks and operations.

Figure 2A. Per-test cost for BD MAX

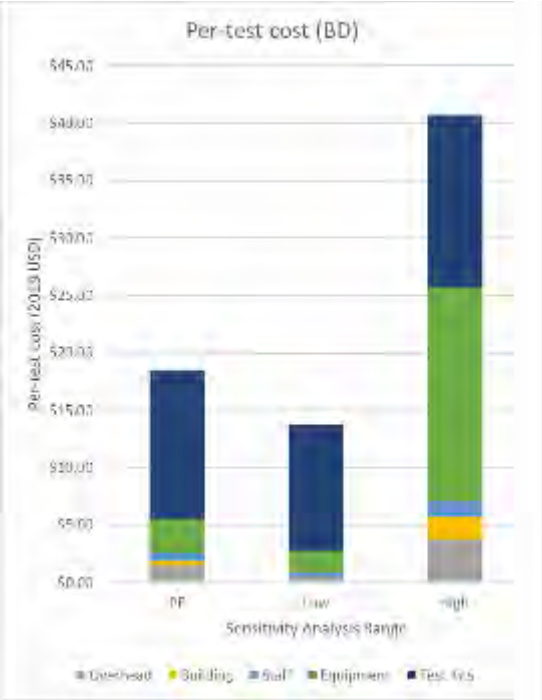
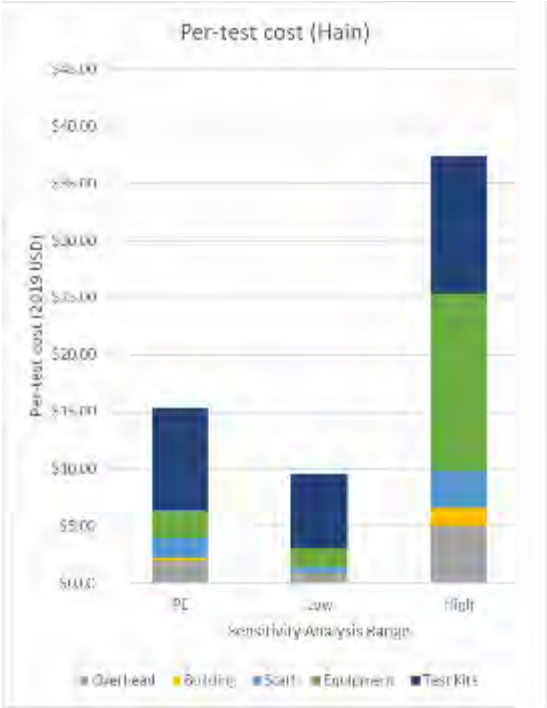


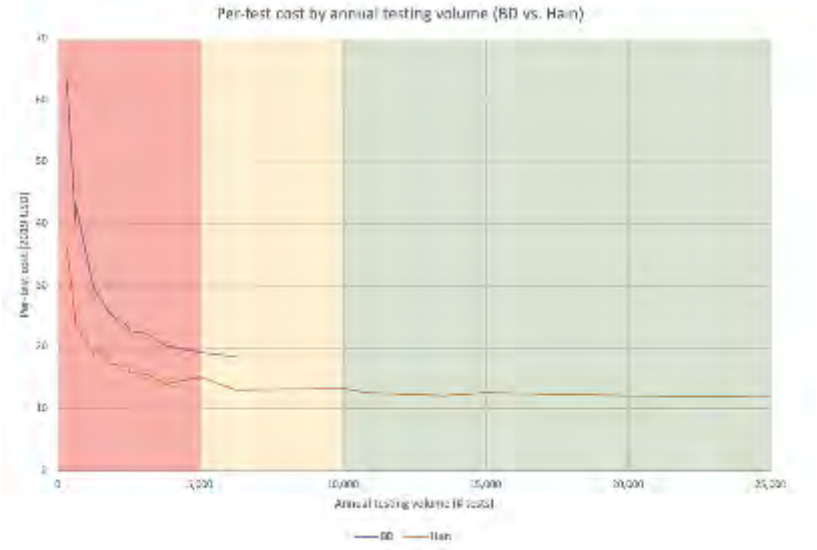
Figure 2B. Per-test cost for Hain



While Hain can run up to 96 samples/run, both platforms were assumed to run only 24 tests/run, equivalent to the capacity of BD MAX 24 samples/run. This resulted in unused capacity for Hain and a likely increase of ~$2/test. Under conditions of full testing capacity for Hain the unit test cost for Hain would be likely even less costly than BD Max.

In one-way sensitivity analyses, annual testing volumes were varied from <5,000 to >25,000 tests/year (Figure 3). Per-test cost was highly sensitive to testing volume when fewer than 5000 tests were conducted per year, but unit test costs begin to stabilize between 5,000 and 10,000 tests /year, and above 10,000 test per year, unit cost estimate was robust. When equipment can be multiplexed and used at capacity, per test cost can be minimized.

Figure 3. One-way sensitivity analysis of per-test cost and annual testing volume.



Additional literature searches conducted to look for economic data using similar platforms from non-TB disease areas identified 3 additional studies from HIV and HCV with very limited cost data, including one study(Boyer et al., 2013) using Abbott RealTime HIV and two on HCV. Data were limited to unit test kit cost and are not transferrable to test kit costs for tests being considered in this review.

How large are the resource requirements (costs)?

Unit test costs for BD MAX and Hain ranged from $18.52 ($13.79 - $40.70) and $15.37 ($9.61 – $37.40), with cheaper per test kit costs reported for Hain and higher operational costs associated with lab processing time. Equipment costs were strong drivers of cost variation and will vary across lab networks and operations, if equipment can be optimally placed or multiplexed to ensure high testing volume, per test cost can be minimized.

What is the certainty of the evidence of resource requirements (costs)?

Available per-test cost data while unpublished, did include overhead, equipment, building, staff and consumable costs however complete quality assessment of the study was not possible. Test cost will vary according to testing volume and laboratory operations. There is limited evidence to assess the important variability across sites, countries and implementation approaches.

Does the cost-effectiveness of the intervention favor the intervention or the comparison?

No studies were identified that assessed cost-effectiveness analyses for any of the E2E solutions and extrapolation was not appropriate given differences in standard of care, different care cascades and associated costs, operational conditions, testing volume and diagnostic accuracy. Implementation considerations such as test placement, lab network, and ability of program to initiate treatment quickly will all likely impact unit test cost and cost-effectiveness. Economic modelling is needed across various settings to understand the range cost-effectiveness profiles of E2E solutions and how they likely vary under different operational criteria.

Cartridge-based nucleic acid amplification tests for detection of XDR-TB (CBNAT) – PICO 2

Two CBNAT tests were identified: the MeltPro® MTB/RIF, Zeesan Biotech Co Ltd. China and the Gene Xpert® MTB/XDR Assay (Xpert XDR, Cepheid, Sunnyvale USA). Only data concerning Xpert MTB/XDR is included in this review. As is the case with Xpert MTB/RIF, the novel XDR assay can be used to test either unprocessed or concentrated sputum. No published studies providing direct evidence on the cost or cost-effectiveness of CBNATs were identified.

Through direct communication from the Xpert XDR manufacturer, Cepheid, the high burden developing country (HBDC) cost for the XDR cartridge is expected to be $19.80USD ex works. Shipping and customs costs will be additional and be borne by the ordering nations or organizations as is the current case for Xpert MTB RIF and Ultra cartridges.

As with Xpert MTB/RIF and Ultra the test cartridge costs represent just one component of the total unit test costs that must be considered, equipment is another important consideration. The Xpert MTB/XDR will not work on existing 6-colour modules and require upgrading to 10-colour GeneXpert modules. Once approval and registration for these new modules is in place in a given country, Cepheid will start only supplying the 10-colour modules. There will be different upgrade options to the 10-colour system with different price points depending on needs and resources available. Upgrade options include 1) a new 10 colour system (most costly option: $9,420 1 module-$72,350 16 modules) including the GeneXpert platform, computer and scanner, 2) a new 10-colour satellite instrument with the GeneXpert connected to an existing system ($6495 1 module-$69,525 for 16 modules) and 3) convert existing GeneXpert system from 6-colour to 10-colour by replacing modules (10 colour module kit $3860). Cepheid flyer on upgrading to the 10-colour modules and associated pricing are included in the supplemental.

Additional cost considerations for Xpert MTB/XDR include additional testing or repeated testing in the case of indeterminate or non-actionable results. The potential cost burden of this will likely vary depending on the proportion of indeterminate test results across settings and the associated re-testing protocols.

No studies assessed cost-effectiveness of the Xpert MTB/XDR cartridge. While extrapolation from other platforms and testing approaches for costing may be appropriate, extrapolation of cost-effectiveness data from Xpert MTB/RIF or other CBNATs is not advised due to differences in diagnostic accuracy, costs associated with XDR treatment and the testing and treatment cascade of care.

Several factors that are likely influential on cost-effectiveness of Xpert MTB/XDR include diagnostic accuracy which may lead to more or fewer individuals being diagnosed compared with the standard of care, and this will vary by region depending on what the standard of care is across different regions. Along with diagnostic accuracy associated with the test itself, the diagnostic algorithm and placement of the Xpert MTB/XDR test within that algorithm has important implications as well.

The novel Xpert MTB/XDR provides results in <90 minutes and therefore introduction of this test will likely result in faster time to result for molecular DST and could impact CEA by improving numbers initiating treatment, reducing loss to follow-up and improving survival rates. Costs associated with XDR treatment are likely an important driver of cost and cost-effectiveness as previous work has shown these costs are large compared to diagnostic and other treatment costs. As larger numbers of XDR positive individuals requiring treatment are identified, total resources required to treat these individuals will increase.

In the absence of transmission modelling studies, little can be said about the long-term population level impact of introducing Xpert MTB/XDR, however the benefits of identifying more cases earlier could lead to a reduction in ongoing transmission and potential cost-savings over longer time horizons. This would need to be thoroughly investigated in a modelling approach that can account for ongoing transmission in the population.

An additional source of indirect evidence given the lack of published studies was extrapolation from costs associated with Xpert MTB/RIF. From an earlier systematic review completed by our group for a December 2019 WHO GDG meeting on *Molecular assays intended as initial tests for the diagnosis of pulmonary and extrapulmonary TB in adults and children*, we concluded that average unit test costs for Xpert MTB/RIF was $34.27, ranging from $11.37 USD-$90.5USD, with higher Xpert unit test costs from China, Germany and the USA. Compared with unit test costs for culture, Xpert MTB/RIF costs may actually be cost-neutral or cost-savings depending on the setting. When considering downstream costs associated with increased numbers of patients diagnosed with TB and MDR-TB through Xpert MTB/RIF, TB treatment costs and ART costs and HIV care costs are expected to increase with increased sensitivity of Xpert MTB/RIF(Abimbola et al., 2012; Adelman et al., 2018; Andrews et al., 2012; S. et al., 2018; Zwerling et al., 2015).

Unit test costs associated with Xpert MTB/RIF varied across settings, and depended on implementation conditions, location of test (centralized versus peripheral or POC settings, test volume and on what cost elements were included in unit test cost estimates (ie. Staff, overhead, training, maintenance, sample transport, etc.)

How large are the resource requirements (costs)?

No direct evidence from published studies regarding total resources required. Resource requirements will include the purchase of cartridges ($19.80USD/cartridge), upgrading of existing platforms to 10-colour modules (an upgrade that will be required eventually for all Xpert platforms: $3860 to >$72,350) and operational and programmatic costs associated with implementing the novel diagnostic. Resource requirements for XDR treatment (drugs, hospital capacity, staff, etc.) likely will also increase with increasing numbers diagnosed. Total costs will vary depending on testing volume and prevalence of XDR in the population. Budget impact will depend on current standard of care and associated resource use.

What is the certainty of the evidence of resource requirements (costs)?

Direct costs related to cartridge and machinery are provided from the manufacturer while several important items related to resource use including staff time, overhead and operational costs associated with implementing Xpert MTB/XDR have not been investigated. Differences in resource use between Xpert MTB/XDR and existing approaches will vary across settings using different phenotypic and genotypic DST. Important variability exists in costs of staff time and operational costs, such as testing volume across settings.

Does the cost-effectiveness of the intervention favor the intervention or the comparison?

No cost-effectiveness studies were identified using XpertMTB/XDR. Extrapolation of cost-effectiveness data from Xpert MTB/RIF or other CBNATs is not advised due to differences in diagnostic accuracy, costs associated with XDR treatment and the testing and treatment cascade of care.

Pyrazinamide (PZA) line probe assays for detection of PZA resistance – PICO 3

No published studies were identified assessing costs or cost-effectiveness using the commercially available PZA LPA test, Genoscholar PZA-TB II, Nipro Japan. The Genoscholar PZA LPA costs $16 USD/test (test kit consumables cost only), with equipment costs for the Multiblot machine at $14,000USD. Nipro has indicated their hopes that further reductions in test cost can be achieved upon global dissemination of the Genoscholar PZA-TB product.

Indirect evidence was available from several sources. One identified study also provided limited costing data on a non-commercial based in-house assay, reporting that costs of test consumables were less than $0.50 excluding costs for DNA extraction and isolate culture(Whitfield et al., 2020). This cost however is likely not comparable with commercially available platforms and test kits.

Four other studies examining other commercially available LPAs (Genotype MTBDRsl and MTBDRplus, Hain Lifesciences) were identified.

In a detailed reference laboratory cost analysis from South Africa, Shah et al assessed laboratory costs for conventional automated liquid culture-based methods, the Xpert MTB/RIF test and the Genotype MTBDRplus line probe assay finding cost per test estimates of $16.88/sample, $14.93 and $23.46/sample respectively in 2013(Shah et al., 2013). Among molecular testing, cost of consumables including cartridge or test kit was by far the largest component cost.

Pooran et al conducted a cost analysis of diagnostic and treatment costs associated with drug resistant TB in South Africa, however diagnostic cost estimates in this study were provided by the National Health Laboratory Service in South Africa(Pooran et al., 2013). Unit test costs were in 2011 USD and ranged from $21.39 for Xpert MTB/RIF to $26.74 for a 1st line DST line probe assay. Detailed costing was performed for other components of diagnosis namely sample transport: $2.72 USD.

Groessl et al compared three rapid tests (MODS, LPA and PSQ) for the diagnosis of XDR-TB with MGIT DST across three countries: Moldova, India, South Africa and found a mean cost/sample of $72.16USD for LPA (MTBDRsl and MTBDRplus) compared with $34.39 for MODS, $50.32 for MGIT, and $59.40 for PSQ(Groessl et al., 2018). Costs included consumables, staff time, equipment and overhead. Cost per

sample for LPA ranged across countries from $33.73 in Moldova to $100.50 in South Africa. LPAs became less expensive than PSQ if testing volume and batch size were maximized. Key drivers of cost included test volume, batch size and staff wages, with South African sites driving higher average test costs due to low volume and high staff wages.

Li et al described the performance and costs of the Genotype MTBDRplus line probe assay (LPA) among inpatients in China, and found conventional DST cost $50.72 per sample, while the LPA unit test cost was estimated to be $108.70(Li et al., 2019). Authors concluded that the total health care costs could be decreased by 71% for smear-positive cases and 25.9% for smear-negative cases, due to shortened turn-around times for the LPA which reduced empirical treatment costs.

In an additional unpublished study from Brazil, de Almeida et al of the Federal University of Minas Gerais & Federal University of Rio de Janeiro found average test cost for Xpert to be $15.60, Genotype MTBDRplus was $76.30 and MGIT $215.75 in 2019USD. Authors concluded MTBDRplus was 4.9 times more costly than Xpert MTB/RIF.

Of note, while the Genoscholar PZA LPA was developed for use with the Nipro automated MultiBlot, a recent unpublished trial demonstrated the Twincubator by Hain could be used successfully, which could improve multiplexing options in some settings.

How large are the resource requirements (costs)?

No direct evidence from published studies regarding total resources required. Resource requirements will include the purchase of test kits (Genoscholar PZA: $16 USD/test kit consumables only), and the equipment which is available for $14,000USD. Operational costs are frequently several fold greater than test kit costs and are not accounted for, and will vary across settings. Unit test costs for Genotype MTBDRsl and MTBDRplus ranged from $23.46 to $108.70, with higher unit test costs coming from settings and countries such as South Africa and China and largely driven by higher staff wages and operational costs. Extrapolations from unit test costs using different LPAs should be done with caution and are not intended to be directly transferrable estimates. These indirect data do suggest that total unit test cost of the Genoscholar PZA LPA is likely several fold higher than unit test kit consumable cost of only $16USD.

Total costs will vary depending on testing volume, numbers eligible for testing and prevalence of PZA resistance in the population. Budget impact will depend on current standard of care, diagnostic and care pathways and associated resource use.

What is the certainty of the evidence of resource requirements (costs)?

Direct costs related to test kits and machinery are available while several important items related to resource use including staff time, overhead and operational costs associated with implementing Genoscholar PZA LPA have not been investigated. Differences in resource use between Genoscholar PZA LPA and existing approaches will vary across settings using different phenotypic and genotypic DST. Important variability exists in costs of staff time and operational costs, such as testing volume across settings.

Does the cost-effectiveness of the intervention favor the intervention or the comparison?

No cost-effectiveness studies were identified using the Genoscholar PZA-TB II. Extrapolation of cost-effectiveness data from other line probe assays is not advised due to differences in diagnostic accuracy, resistance prevalence, and the testing and treatment cascade of care.

PRINCIPAL FINDINGS

- Operational costs likely a large proportion of unit test costs and are largely unknown aside from equipment. These costs are also not always included in unit test costs making comparisons across estimates, methods and studies difficult.
- Batch size and test volume are important drivers of unit test cost and cost-effectiveness.
- Personnel costs vary across regions and are important drivers of cost variation
- Implementation considerations (test placement, lab networks, sample transport costs, ability of program to initiate treatment quickly, etc.) will impact unit test costs and/or cost-effectiveness and limit extrapolation.
- When equipment can be multiplexed and used at capacity, per test cost can be minimized and cost-effectiveness may improve.
- Should aim for standardization of costing approaches and methods to ensure comparability between studies.
- Cost-effectiveness data is needed across various settings and will vary widely with different operational criteria.

DISCUSSION

This systematic review revealed no published literature providing direct evidence and only limited evidence from unpublished sources or from indirect evidence using different diagnostic tests.

Test unit cost estimates are available for some test classes but due to a variety of methods and approaches across studies different component costs are often included; as a result, unit test costs may not be comparable or generalizable. For example, operational costs are often several fold higher than test kit cost yet may not always be included in test unit costs. Staff costs are not always included and can lead to large variations across sites due to differences in salary scales between settings.

Batch size and test volume were consistently important drivers of cost and cost-effectiveness, therefore placement of the test within the laboratory network will be critical. Networks that can maximize testing and multiplexing can reduce unit test costs and improve cost-effectiveness.

Understanding total resource requirements or total costs of a novel program depend on a number of setting specific factors including the number of individuals who will be tested with the novel diagnostic, prevalence of resistance in the tested population, and the current standard of care using phenotypic or genotypic DST (for budget impact assessment). Important considerations for budget impact also include

the additional cost of treating more MDR and XDR patients in scenarios where sensitivity is improved, and more patients are initiating anti-TB treatment.

Cost-effectiveness studies providing direct evidence were not identified and extrapolation is not advised. Cost-effectiveness may be impacted by changes in diagnostic accuracy, differences in standard of care across sites, diagnostic algorithm, reductions in turn-around-time among others. Without proper economic modelling, potential cost-effectiveness of various tests and testing approaches remains unknown.

CONCLUSION

Very limited direct evidence was identified**.** Implementation considerations will vary widely across settings and this variability will impact unit test costs and cost-effectiveness. Cost-effectiveness and transmission modelling data is needed across a variety of scenarios and populations to understand the population level impact and cost-effectiveness implications of introducing these novel diagnostics.

REFERENCES

Abimbola, T. O., Marston, B. J., Date, A. A., Blandford, J. M., Sangrujee, N., & Wiktor, S. Z. (2012). Cost-effectiveness of tuberculosis diagnostic strategies to reduce early mortality among persons with advanced HIV infection initiating antiretroviral therapy. *Journal of Acquired Immune Deficiency Syndromes*, *60*(1), 1–7. https://doi.org/10.1097/QAI.0b013e318246538f

Adelman, M. W., McFarland, D. A., Tsegaye, M., Aseffa, A., Kempker, R. R., & Blumberg, H. M. (2018). Cost-effectiveness of WHO-Recommended Algorithms for TB Case Finding at Ethiopian HIV Clinics. *Open Forum Infectious Diseases*, *5*(1), 1–8. https://doi.org/10.1093/ofid/ofx269

Andrews, J. R., Lawn, S. D., Rusu, C., Wood, R., Noubary, F., Bender, M. A., Horsburgh, C. R., Losina, E., Freedberg, K. A., & Walensky, R. P. (2012). The cost-effectiveness of routine tuberculosis screening with Xpert MTB/RIF prior to initiation of antiretroviral therapy: A model-based analysis. *Aids*, *26*(8), 987–995. https://doi.org/10.1097/QAD.0b013e3283522d47

Boyer, S., March, L., Kouanfack, C., Laborde-Balen, G., Marino, P., Aghokeng, A. F., Mpoudi-Ngole, E., Koulla-Shiro, S., Delaporte, E., Carrieri, M. P., Spire, B., Laurent, C., & Moatti, J. P. (2013). Monitoring of HIV viral load, CD4 cell count, and clinical assessment versus clinical monitoring alone for antiretroviral therapy in low-resource settings (Stratall ANRS 12110/ESTHER): A cost-effectiveness analysis. *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(13)70073-2

Groessl, E. J., Ganiats, T. G., Hillery, N., Trollip, A., Jackson, R. L., Catanzaro, D. G., Rodwell, T. C., Garfein, R. S., Rodrigues, C., Crudu, V., Victor, T. C., & Catanzaro, A. (2018). Cost analysis of rapid diagnostics for drug-resistant tuberculosis. *BMC Infectious Diseases*, *18*(1). https://doi.org/10.1186/s12879-018-3013-0

Li, X., Deng, Y., Wang, J., Jing, H., Shu, W., Qin, J., Pang, Y., & Ma, X. (2019). Rapid diagnosis of multidrug-resistant tuberculosis impacts expenditures prior to appropriate treatment: A performance and diagnostic cost analysis. *Infection and Drug Resistance*, *12*, 3549–3555. https://doi.org/10.2147/IDR.S224518

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. In *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1000100

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. In *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1000097

Pooran, A., Pieterson, E., Davids, M., Theron, G., & Dheda, K. (2013). What is the Cost of Diagnosis and Management of Drug Resistant Tuberculosis in South Africa? *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0054587

S., O., I., T., F., C., I., P., L., P., M.C., M., M., G., M., F., S., M., E., M., & G., T. (2018). Delayed diagnosis and treatment of tuberculosis in HIV+ patients in Mozambique: A cost-effectiveness analysis of screening protocols based on four symptom screening, smear microscopy, urine LAM test and Xpert MTB/RIF. *PLoS ONE*, *13*(7), 1–16. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L623078580%0Ahttp://dx.doi.org/10.1371/journal.pone.0200523

Shah, M., Chihota, V., Coetzee, G., Churchyard, G., & Dorman, S. E. (2013). Comparison of laboratory costs of rapid molecular tests and conventional diagnostics for detection of tuberculosis and drug-resistant tuberculosis in South Africa. *BMC Infectious Diseases*, *13*(1).

https://doi.org/10.1186/1471-2334-13-352

Whitfield, M. G., Marras, S. A. E., Warren, R. M., Van Rie, A., Rice, J., Wangh, L. J., & Kreiswirth, B. N. (2020). Rapid Pyrazinamide Drug Susceptibility Testing using a Closed-Tube PCR Assay of the Entire pncA gene. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-61286-7

Zwerling, A. A., Sahu, M., Ngwira, L. G., Khundi, M., Harawa, T., Corbett, E. L., Chaisson, R. E., & Dowdy, D. W. (2015). Screening for tuberculosis among adults newly diagnosed with HIV in Sub-Saharan Africa: A cost-effectiveness analysis. *Journal of Acquired Immune Deficiency Syndromes*, *70*(1), 83–90. https://doi.org/10.1097/QAI.0000000000000712

SUPPLEMENT

Cepheid 10-Colour Multiplexing technology Upgrade Flyer with costs

10-Color Multiplexing Technology

GeneXpert® System Upgrade Options for HBDC Customers

Choose one of these upgrade options:

1 Convert your current GeneXpert System from 6-color to 10-color modules

Required and must be ordered separately:
One Software kit Dx 6.2 or higher per site for Xpert MTB/XDR testing (supplied free of charge)

- Remote installation included and onsite installation available by request*
- All modules of the same system should be converted (hybrid systems of 6- and 10-color modules are not supported), i.e. for a GeneXpert IV, 4 new modules should be ordered.
- There is no swap or discount option for the existing 6-color modules.

Computer check:

- Computers running WIN97 operating systems must be replaced with a new WIN10 computer.
- Cepheid recommends replacing computers running WIN7 with new WIN10 computers, due to Microsoft's discontinuation of WIN7 support in 2021.

2 Daisy chain a 10-color satellite instrument to your current GeneXpert System

Required and must be ordered separately:
DAISYKIT (§75) and one Software kit Dx 6.2 or higher per site for Xpert MTB/XDR testing (supplied free of charge)

- Remote installation included and onsite installation available by request*

Computer check:

- Computers running WinXP operating systems must be replaced with a new WIN10 computer.
- Cepheid recommends replacing computers running WIN7 with new WIN10 computers, due to Microsoft's discontinuation of WIN7 support in 2021.

3 Order a new GeneXpert System with 10-Color Technology

Includes:
WIN10 Computer, Barcode Scanner, latest GeneXpert Software and all accessories.

- Remote installation included and onsite installation available by request*
- The new 10-color system can also be connected to an existing GeneXpert System and replace an old computer. If this is the case, make sure to add a separate DAISYKIT part to your new system order.

Ordering Information:

Catalog Number(s)	Description	Price
New Systems with Desktop		
GXE-1-D-10C	GeneXpert® II R2, 1-Module, Desktop, 10-color	$9,490
GXE-2-D-10C	GeneXpert II R2, 2 Modules, Desktop, 10-color	$12,480
GXIV-2-D-10C	GeneXpert® IV R2, 2-Modules, Desktop, 10-color	$12,780
GXIV-4-D-10C	GeneXpert IV R2, 4 Modules, Desktop, 10-color	$19,600
GXXVI-8-D-10C	GeneXpert® XVI R2, 8-Modules, Desktop, 10-color	$40,000
GXXVI-16-D-10C	GeneXpert XVI R2, 16-Modules, Desktop, 10-color	$71,650
New Systems with Laptop		
GXE-1-L-10C	GeneXpert II R2, 1 Module, Laptop, 10-color	$9,990
GXE-2-L-10C	GeneXpert II R2, 2-Modules, Laptop, 10-color	$13,040
GXIV-2-L-10C	GeneXpert IV R2, 2-Modules, Laptop, 10-color	$13,390
GXIV-4-L-10C	GeneXpert IV R2, 4-Modules, Laptop, 10-color	$19,500
GXXVI-8-L-10C	GeneXpert XVI R2, 8-Modules, Laptop, 10-color	$40,500
GXXVI-16-L-10C	GeneXpert XVI R2, 16-Modules, Laptop, 10-color	$72,350

Catalog Number(s)	Description	Price
Satellite instruments – DAISYKIT required		
900-0799	GeneXpert I R2, 1-Module, 10-color	$9,490
900-0794	GeneXpert II R2, 2 Modules, 10-color	$12,480
900-0800	GeneXpert IV R2, 2-Modules, 10-color	$12,780
900-0802	GeneXpert IV R2, 4 Modules, 10-color	$19,600
900-0801	GeneXpert XVI R2, 8-Modules, 10-color	$40,000
900-0806	GeneXpert XVI R2, 16-Modules, 10-color	$71,650
DAISYKIT	GeneXpert® Daisy Chain Accessory Kit	$75
New Module		
900-0807	10-color GeneXpert®, Module Kit	$3,690
Accessories		
GXSWKIT	GeneXpert® Dx 6.3 Software	$0
900-0469-GXDX	Win 10 Desktop	$1,600
900-0527-GXDX	Win 10 Laptop	$2,995
900-0220	Monitor, 19", Flat Panel	$268.17

Web Annex D.10. User perspectives on NAATs to detect TB and resistance to anti-TB agents: results from qualitative evidence synthesis (systematic review)

Nora Engel[1], Eleanor A Ochodo[2,3], Perpetua Wanjiku Karanja[4], Bey-Marrié Schmidt[5], Ricky Janssen[1], Karen R Steingart[6], Sandy Oliver,[7,8]

Affiliations:
[1]Department of Health, Ethics and Society, Maastricht University, Maastricht, Netherlands (Kingdom of the)
[2]Centre for Evidence-based Health Care, Department of Global Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[3]Centre for Global Health Research, Kenya Medical Research Institute, Kenya
[4] Kirinyaga County, Department of Health, Kenya
[5]Cochrane South Africa, South African Medical Research Council, Cape Town, South Africa
[6]Honorary Research Fellow, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK
[7]EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London, London, UK
[8]Africa Centre for Evidence, Faculty of Humanities, University of Johannesburg

Produced in preparation for the WHO guideline group meeting on "Nucleic acid amplification tests to detect tuberculosis and drug-resistant tuberculosis" on 7-18 December, 2020.
Please send comments and feedback to Nora Engel, n.engel@maastrichtuniversity.nl

Background

Description of the topic

Tuberculosis (TB) causes 10 million cases and 1.5 million deaths annually and it is estimated that 3 million cases go undiagnosed each year (World Health Organization, 2020a). Drug-resistant TB (DR-TB) is a major threat to global TB control. Ending the global TB epidemic will be achievable over the next 20 years only if there is intensive action by all countries which have endorsed the End TB Strategy and its ambitious targets (World Health, 2015). Early diagnosis and prompt treatment of all persons of all ages with any form of drug-susceptible or drug-resistant TB is fundamental. Until 2018, all multidrug-resistant (MDR)-TB regimens employed at least five second-line drugs for a duration of up to 24 months. The arrival of novel or repurposed drugs such as bedaquiline, clofazimine and linezolid, has revolutionized the efficacy of longer regimens, dispensing with the need for injectable drugs and promising to deliver shorter all-oral regimens (World Health Organization, 2020c). Recently, a six-month three drug regimen based on bedaquiline, linezolid and the novel drug pretomanid achieved high rates of treatment success in an observational cohort of XDR-TB patients (Conradie et al., 2020). Early recognition and characterization of resistance is a prerequisite for effective delivery of these new treatment strategies for DR-TB as quickly as possible to those who could benefit. This draws attention to the need for faster, cheaper and more easily deployable technologies and testing programmes.
WHO-endorsed rapid TB diagnostics and drug susceptibility testing (DST) should be available to all persons with signs and symptoms of TB to meet the targets of the End TB Strategy. Yet, TB diagnosis is a crucial bottleneck in many countries. While the availability of drug susceptibility testing using culture-based and molecular methods is increasing, coverage and availability of these technologies varies widely. For example, globally in 2019, only 59% of bacteriologically confirmed new tuberculosis cases were tested for rifampicin resistance (World Health Organization, 2020a).

The diagnosis of tuberculosis and drug-resistant forms has seen important changes and innovations over the last years. One of these has been the introduction of automated nucleic acid amplification tests (NAATs) of low-complexity, designed to work outside well-equipped, often centralized, laboratories that are difficult to access for most patients. NAATs are molecular systems that can detect small quantities of genetic material (DNA or RNA) from microorganisms, such as *Mycobacterium tuberculosis*, by amplifying the quantities to an amount large enough for studying in detail. A variety of molecular amplification methods are available, of which polymerase chain reaction (PCR) is the most common. This review focuses on low-complexity NAATs. Low complexity refers to a situation where no special infrastructure is required and basic laboratory skills are suitable to run the test. However, equipment may still be required. A presumed key advantage of NAATs is that they are rapid diagnostic tests, potentially providing results in a few hours. This is particularly promising for tuberculosis, where diagnostic and treatment delays are often substantial. Diagnostic devices only have an impact if they are put to use in a correct and timely manner. In the case of automated NAATs of low complexity for TB and drug-resistant forms of TB, this involves patients and patient contacts who seek care, produce a sample and return for results; healthcare workers who order, conduct the diagnostic and then act on the result; healthcare workers and technicians or suppliers who order stock and maintain the machines, but also programme officers who deploy and monitor these devices. These users matter in ensuring the functioning and utilization and therefore the impact the diagnostic can have. It is essential to understand the perspectives and experiences of these users with automated NAATs of low complexity to inform policy, funding, research and development.

How this review might inform or supplement what is already known in this area

Qualitative evidence on user perspectives has only recently been commissioned as stand-alone primary studies for specific technologies (World Health Organization, 2020b) but has never been systematically

reviewed for a group of technologies. Current WHO guidance on automated NAATs of low complexity for TB diagnosis is based on systematic reviews of diagnostic accuracy and cost-effectiveness. Accuracy studies do not reveal what users think of or experience with the diagnostic in question. Yet to understand why diagnostics are utilized, how effective they are and their impact on health equity, it is essential to answer questions around feasibility, added value and experiences which our review findings will provide.

How the intervention might work

The promise of automated NAATs of low complexity for tuberculosis and drug-resistant forms of tuberculosis is that they can be done closer to where tuberculosis patients are, in more peripheral settings of the community and thereby cut diagnostic delay, provide a more accurate diagnosis of tuberculosis and a diagnosis of drug resistance and thereby have important implications for patient important outcomes (Bainomugisa, Gilpin, Coulter, & Marais, 2020; Pooran et al., 2019). While there is no clear statistical evidence of a significant effect of Xpert MTB/RIF on all-cause mortality (Di Tanna et al., 2019; Haraka & al., 2019)., early detection of tuberculosis and rifampicin resistance may not lead to improved patient outcomes if the test result is not linked to appropriate treatment and other healthcare services (Pai, Schumacher, & Abimbola, 2018).
Our review will not consider the effectiveness of NAATs or their quantifiable impact on patient important outcomes. Rather, we are concerned with the perspectives and experiences of various users in dealing with these technologies in their work and routines.

Why is it important to do this review?

The users of diagnostics include patients and their contacts, clinic staff, laboratory managers, ministries of health, and implementers. If we do not take the perspective of all users into consideration, we risk that these technologies do not fit their intended setting of use, cannot be made to work and scaled up and are not utilized or not accessible for those in need. User perspectives on new diagnostics, their preferences and values, acceptability, and feasibility are important to take into account during WHO decision-making on new diagnostics and guideline development.

Challenges with implementation and underutilization
The United Nations Sustainable Development Goals represent a collective plan to end poverty, decrease inequality, and protect the planet from degradation by 2030 (United Nations Sustainable Development Goals 2030). Ending the tuberculosis epidemic by 2030 is among the health-related targets described in United Nations Sustainable Development Goal 3 (WHO End TB 2015). Automated NAATs of low complexity for tuberculosis drug resistance have had an immense influence on tuberculosis policy and care in high burden settings, but there are persistent concerns about underutilization. The proposed review will contribute to reaching SDG3 by ensuring user perspectives including end-user preferences, values and perspectives on feasibility of these diagnostics are being considered in a systematic way during the above WHO decision-making meeting.
Alignment with WHO priorities
This qualitative review complements a Cochrane diagnostic test accuracy review in progress on 'Xpert MTB/XDR for detection of pulmonary tuberculosis and resistance to isoniazid, fluoroquinolones, ethionamide, and amikacin'. These reviews informed the WHO Guideline Development Group Meeting, "Nucleic acid amplification tests to detect tuberculosis and drug resistant tuberculosis" on 7-18 December, 2020.
The three classes of technologies that were evaluated include:

- Index test 1: Low-complexity automated NAATs for detection of resistance to isoniazid and second-line anti-TB agents;

- Index test 2: Moderate-complexity automated NAATs for detection of TB and resistance to rifampicin and isoniazid;
- Index test 3: High-complexity hybridization-based NAATs for detection of resistance to pyrazinamide.

A qualitative evidence synthesis can add value by providing decision makers with additional evidence to improve understanding of intervention complexity, contextual variations, implementation, and stakeholder preferences and experiences.

Objectives

To synthesize user perspectives and experiences with automated NAATs of low complexity for detection of tuberculosis and tuberculosis drug resistance.

Review question

What are the perspectives and experiences of those using automated NAATs of low complexity to diagnose tuberculosis and tuberculosis drug resistance?
Answering this question will allow us to identify the implications for effective implementation and health equity.

Methods

Criteria for considering studies for this review

Types of studies
We included primary studies that use qualitative methodologies or study designs such as ethnography, phenomenology, case studies, grounded theory studies and qualitative process evaluations. We included studies that use both qualitative methods for data collection (e.g. focus group discussions, individual interviews, observation, diaries, document analysis, open-ended survey questions) and qualitative methods for data analysis (e.g. thematic analysis, framework analysis, grounded theory, narrative analysis). We excluded studies that collect data using qualitative methods but do not analyse these data using qualitative analysis methods (e.g. open-ended survey questions where the response data are analysed using descriptive statistics only), because such studies rarely offer the conceptual or contextual detail for understanding the complexities of interventions and their implementation, how these vary with context, or users' perspectives or experiences (Noyes et al., 2020).
We included mixed methods studies where it was possible to extract the data that were collected and analysed using qualitative methods.
We included both published and unpublished studies and studies published in any language.
We included studies regardless of whether they were conducted alongside studies of the effectiveness of automated NAATs of low complexity to tuberculosis and drug resistant forms of TB or independently.
We did not exclude studies based on our assessment of methodological limitations. We used this information about methodological limitations to assess our confidence in the review findings.

Topic of interest
Any qualitative study related to the application of automated NAATs of low complexity for tuberculosis and tuberculosis drug resistance, pathways from diagnosis to treatment including automated NAATs of low complexity, intervention studies, operational research, feasibility and acceptability assessments.
Participants

This review focuses on users and potential users of automated NAATs of low complexity. Users include patients and their caregivers, laboratory technicians, healthcare providers, implementers and programme officers who are involved in diagnosing and treating tuberculosis and drug resistant forms of TB as well as ordering, operating, maintaining diagnostics and acting on diagnostic test results. Potential users include users who do not (yet) utilize the diagnostic for instance because they are unable to access it or make it work within their routines or setting.

Setting

We included studies on automated NAATs of low complexity located in any country, including low-, middle- and high-income countries located in any setting, including centralized and more peripheral locations in a health system and any type of health facility (hospital, peripheral laboratory, clinic, community health centre or mobile testing van).

Health issues

Participants of qualitative studies may or may not have symptoms of any respiratory illness.

Intervention

Diagnostic testing that involves automated NAATs of low complexity for example, but not limited to, Xpert MTB/RIF, Xpert MTB/XDR, and Truenat.

Search methods for identification of studies

We developed the search strategy in collaboration with the Information Specialist from the Cochrane Infectious Diseases Group. We also consulted the EPOC Information Specialist before developing the strategy. We attempted to identify all relevant studies regardless of language or publication status (published, unpublished, in press, and in progress).

Electronic searches

We searched the following databases from 1t January 2007 to 5th September 2020, using the search terms and strategy described in Appendix 1:

- Medline (OVID);
- Embase (OVID);
- CINAHL (EBSCOHost; Cumulative Index to Nursing and Allied Health Literature);
- PsycInfo (EBSCOHost).

Searching other resources

We contacted researchers and experts in the field to identify any additional eligible studies. We also checked the references of relevant reviews and studies to identify additional studies.

Grey literature

Due to time and resource constraints, we did not conduct an extensive grey literature search. We did ask members of the GDG and within our personal networks for unpublished reports of implementing partners and technical agencies.

Selection of studies

We used Covidence to manage the selection of studies (Covidence, 2017). Two review authors independently scrutinized titles and abstracts identified from literature searching to identify potentially eligible studies. We retrieved for full text review the article of any citation identified by one of the review authors as potentially eligible. We then as a team rescreened the preliminary included studies and excluded those that did not meet the full text screening conditions (e.g. conference abstracts or review article instead of primary study). Then, two review authors independently assessed full-text articles for inclusion using predefined inclusion and exclusion criteria. For both screening steps, we resolved disagreements by discussion, if necessary, with a third review author. We recorded all studies excluded after full-text

assessment and their reasons for exclusion in the Characteristics of excluded studies table (appendix 5). We illustrated the study selection process in a PRISMA diagram (Moher, Liberati, Tetzlaff, & Altman, 2009).

Language translation

We searched for primary studies irrespective of their language of publication.

Sampling of studies

This qualitative evidence synthesis aims to describe the experiences of those using automated NAATs of low complexity for diagnosis of tuberculosis and DR-TB in a coherent way. After we identified all studies that were eligible for inclusion, we assessed the number of studies and the data richness or thickness available for synthesis. Because we found a rather large number of studies that met our inclusion criteria (27), purposefully selected a first sample of eligible studies with rich or thick data and a second sample of other studies that addressed various users, uses of and experiences with the intervention not addressed by the richer/ thicker studies. To do so we first categorized the eligible studies into rich or thick and poor or thin studies depending on the depth of the analysis undertaken. A thick study is one in which the author 1) analyzed their findings beyond a descriptive list of barriers/facilitators, 2) demonstrates insights into participants perspectives and experiences, 3) portrays richness and complexity of the data (i.e. explains variation and illustrates meanings, 4) develops or contributes to theory (this approach has been used in (Eshun-Wilson, Rohwer, Hendricks, Oliver, & Garner, 2019)). This generated six studies with very rich or thick data, and six studies with very poor or thin data. The remaining 15 studies had data of medium richness/thickness. We decided to take forward for review the six very rich/thick studies and 15 medium rich/thick studies as this final sample also covered variations in users (different healthcare providers, patients, decision-makers), uses of (location of testing site, role in diagnostic algorithm) and experiences with the intervention (positive, negative).

Data extraction

Five review authors (EO, NE, BS, PW, RJ) extracted the following data from eligible studies:

- Descriptive study-related information: Study author, year of publication, language, study location (country, rural/urban, public/private, type of facilities), background prevalence of (MDR-)TB;
- Study objectives and rationale, method of data collection, method of data analysis, conceptual framework if used, how the study was conceived (independence of those designing, implementing or evaluating the intervention);
- Intervention-related information: the type of (potential) user involved (e.g. patients, clinicians, nurses, laboratory staff, implementers), the diagnostic tools used, programmatic features of the intervention (e.g. testing model/algorithm/program in which the diagnostic is used, including the target population, setting and eligibility criteria, the envisioned role of the diagnostic (e.g. replacement, add-on),sample transport and result communication);
- Key study findings were extracted in narrative form in MS Word (for instance, qualitative themes/categories/findings/supporting quotations, and conclusions, the type and rate of use emerging from the study findings (e.g. batching, number of tests run on average, underutilization). Among the key findings we also extracted data (if available) on the following factors that, based on our prior research experience, we expected to be important to user experiences: added value to the particular user, workflow, resources involved in implementing it, confidence in test results, implementation process, access/equity.

Two review authors individually extracted data from the same study and resolved any conflicts in a consensus meeting. Authors of primary studies did not extract data from their own study or studies. Instead, another author extracted these data.

Assessing the methodological limitations of included studies

Two review authors [any pair from NE, BS, PW, EO] independently assessed methodological limitations for each study using the EPPI-Centre tool (Rees, Caird, Dickson, Vigurs, & Thomas, 2014). We resolved disagreements by discussion. Team members who are authors of included studies did not assess the methodological limitations of their own studies. We assessed methodological limitations according to the following domains:

- Rigor in sampling: -the sampling strategy was appropriate to the questions posed in the study (e.g. was the strategy well-reasoned and justified?); -attempts were made to obtain a diverse sample of the population in question (think about who might have been excluded; who may have had a different perspective to offer); -characteristics of the sample critical to the understanding of the study context and findings were presented (i.e. do we know who the participants were in terms of, for example, basic socio-demographics, characteristics relevant to the context of the study, etc.).
- Rigor in data collection: -data collection tools were piloted; -data collection was comprehensive, flexible and/or sensitive enough to provide a complete and/or vivid and rich description of people's perspectives and experiences (e.g. did the researchers spend sufficient time at the site/with participants? Did they keep 'following up'? Was more than one method of data collection used?); - steps were taken to ensure that all participants were able and willing to contribute (e.g. processes for consent, language barriers, power relations between adults and children/young people).
- Rigor in data analysis: -data analysis methods were systematic (e.g. was a method described/can a method be discerned?); -diversity in perspective was explored; -(if qualitative) the analysis was balanced in the extent to which it was guided by preconceptions or by the data); -the analysis sought to rule out alternative explanations for findings (in qualitative research this could be done by, for example, searching for negative cases/exceptions, feeding back preliminary results to participants, asking a colleague to review the data, or reflexivity; in quantitative research this may be done by, for example, significance testing).
- Extent to which findings are grounded in/supported by the data:-enough data are presented to show how the authors arrived at their findings; -the data presented fit the interpretation/support claims about patterns in data; -the data presented illuminate/illustrate the findings; -(for qualitative studies) quotes are numbered or otherwise identified and the reader can see that they don't just come from one or two people
- Breadth and depth of findings: Consider whether: -a range of issues are covered; -the perspectives of participants are fully explored in terms of breadth (contrast of two or more perspectives) and depth (insight into a single perspective); -richness and complexity has been portrayed (e.g. variation explained, meanings illuminated); -there has been theoretical/conceptual development.

We reported our assessments in a Methodological Limitations table (see Appendix 2).

Data management, analysis and synthesis

We used a thematic approach to guide data analysis. We synthesized qualitative research to better understand views and experiences with the intervention in context of use. From this understanding we deduced values, feasibility and acceptability considerations for low complexity automated NAATs for TB and drug resistant TB.

Based on the key findings extracted from the six rich studies, NE developed a coding scheme which was discussed with the other review authors. Using the coding scheme, NE coded the key findings of the 6 studies with rich data using NVIVO (version 12) and wrote memos on selected themes. In a second round of analysis, data from the 15 studies of medium richness was extracted by NE, EO, PW and BS; and NE coded these summaries in the same way as the 6 rich studies. NE then added the emerging additional insights and data to the existing memos. In a next step, NE generated review findings based on these memos, which were revised and finalized after discussion with the other review authors. Finally, we developed a figure to illustrate how our findings hang together (see Figure 2 below).

Assessing our confidence in the review findings

Two review authors [NE, EO in consultation with BS] used the GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative research) approach to assess our confidence in each finding (Lewin et al., 2015). CERQual assesses confidence in the evidence, based on the following four key components:
1. Methodological limitations of included studies: the extent to which there are concerns about the design or conduct of the primary studies that contributed evidence to an individual review finding.
2. Coherence of the review finding: an assessment of how clear and cogent the fit is between the data from the primary studies and a review finding that synthesizes those data. By cogent, we mean well supported or compelling.
3. Adequacy of the data contributing to a review finding: an overall determination of the degree of richness and quantity of data supporting a review finding.
4. Relevance of the included studies to the review question: the extent to which the body of evidence from the primary studies supporting a review finding is applicable to the context (perspective or population, phenomenon of interest, setting) specified in the review question.
After assessing each of the four components, we made a judgement about the overall confidence in the evidence supporting the review finding. We judged confidence as high, moderate, low, or very low. The final assessment is based on consensus among the review authors. All findings start as high confidence and are then graded down if there are important concerns regarding any of the CERQual components.
Because the criteria 'Breadth and depth of findings' of the EPPI-Centre tool and the component 'adequacy' of CERQual overlap and to avoid double counting problems, we did not use the information on breadth and depth of findings in our assessment of 'methodological limitations' but only for assessing 'adequacy'.

Summary of Qualitative Findings table(s) and Evidence Profile(s)

The summary of qualitative findings table with detailed descriptions of our confidence assessment is available in appendix 3.

Review author reflexivity

The author team represents a diversity in disciplinary backgrounds, research foci and experiences with both qualitative and quantitative study designs for both, primary empirical research and evidence synthesis. Together, they have experience with diverse fields of study (public health (RJ, SO, EO, KS, BS), Science and Technology Studies (NE, RJ), medical sociology and anthropology (NE, BS, RJ), epidemiology (EO), health systems (SO), qualitative synthesis methodology (SO), pharmacoepidemiology, pharmacovigilance (PW)), experience with different geographical settings and experience with researching diagnostic processes and technologies (ranging from technical accuracy studies to studies of healthcare seeking, implementation challenges, point-of-care testing processes and evaluation of specific diagnostic devices). Such a multidisciplinary team facilitates analysis and identification of multiple factors influencing user perspectives and feasibility considerations. At the outset of the review, some of the authors would anticipate that automated NAATs of low complexity have the potential to improve TB care, but that critical barriers

exist to their implementation, while others might be more hesitant about the presumed automatic benefit of introducing advanced technologies but then not investing in strengthening weak health systems, while again some might wonder about the design process of automated NAATs of low complexity and to what extent it took into account perspectives of various users. All would have been in contact with different types of users throughout their research career. We minimized the risk that our perspectives as authors influence the analysis and interpretation, by using refutational analysis techniques, such as taking seriously contradictory findings between studies and further exploring and analysing them. We used the different perspectives represented in the author team productively in regular meetings wherein we discussed emerging findings and themes with the aim of identifying our underlying assumptions in the data synthesis, clarifying procedures and documenting various challenges faced in the review process. This supported and enhanced the reflexivity of the review team. We described these differences and issues contributing to the interpretation of the review findings in the reflexivity section in the full manuscript.

NE has conducted a range of primary studies in India and South Africa's health system examining challenges to diagnosing and diagnostic processes at point of care. She has also undertaken studies on the attempts of innovating and implementing point-of-care diagnostics for TB and HIV, among them automated NAATs of low complexity. She uses a constructivist viewpoint/epistemology that is sensitive to how technology design and use mutually constitute each other, meaning that users are influenced by and also shape technologies, not only once technologies are developed and in use, but also when assumptions about users are inscribed into material characteristics of technologies such as automated NAATs of low complexity. These prior experiences might make her particularly sensitive to challenges in implementation and the perspectives of a wide variety of users.

EO is a public health physician and methodologist. She has 10 years experience in evidence synthesis specializing in methodology, systematic reviews and meta-analysis of diagnostic tests. She has conducted systematic reviews on TB tests, some of which have informed WHO guidelines on TB tests. Eleanor is also an academic editor with Cochrane Infectious Disease Group.

PW has no prior experience with TB diagnostics research. Her views on TB diagnostics are primarily influenced by being a health care worker involved in a multidisciplinary review of MDR TB patients management.

BS is a public health researcher with experience in conducting qualitative and quantitative, Cochrane and non-Cochrane systematic reviews. She has conducted some primary research on TB-related topics previously. Her systematic review expertise were valuable in guiding the review team with specific processes, specifically in terms of data extraction and analysis and assessing the confidence in review findings.

RJ has minimal experience in the field of tuberculosis diagnostics. She has conducted qualitative research regarding the implementation of digital strategies for HIV self-testing and HIV testing at point-of-care in South Africa. She also has a background in biological sciences and some practical and theoretical knowledge regarding basic laboratory methodology. These experiences make her sensitive to the importance of valuing new diagnostics for their accuracy and reliability within the laboratory, but also the necessity of implementing new diagnostics such that the information they provide can be applied in clinical practice to enable good patient care.

KRS is a public health physician and methodologist. She has performed over 20 systematic reviews on TB tests and contributed to several recent WHO policies on TB diagnostics. Karen is an Editor, with Cochrane Infectious Disease Group and Cochrane Diagnostic Test Accuracy Editorial Team.

SO has no personal experience regarding TB diagnostics and began this work agnostic about automated NAATs of low complexity. She views interventions primarily from the standpoint of patients, families and the wider public. She has been systematically reviewing research about program effectiveness and implementation, and experiences of the providers and potential recipients, for 25 years. She is an editor with the Cochrane Consumers and Communication Review Group and the Cochrane Infectious Disease Group.

Results

Results of the search

We found 27 studies that met our inclusion criteria. We sampled 21 of these studies for inclusion in the analysis (see Figure 1: Flow chart Appendix 4). All of the sampled studies were published between 2012 and 2020. Six studies remain to be reviewed which were not sampled as we assessed them as particularly thin (see section sampling). They do not address participants not already addressed in the sampled studies. For an overview of the studies that were not sampled and studies that were excluded, see Appendix 5.

Description of the studies

A summary of the core characteristics of studies included in this review is presented in a study characteristics table in appendix 6. Of included studies, all were located located in high TB burden countries with 4 in South Africa, 1 in Vietnam, 6 in India, 1 in Nepal, 2 in Uganda, 1 in Brazil, 2 in Kenya, 1 in Eswatini, 1 in Myanmar, 1 in Mongolia and 1 in Cambodia. In addition, one study covered projects in nine countries (Democratic Republic of Congo (DRC), Kenya, Pakistan, Bangladesh, Mozambique, Cambodia, Malawi, Nepal, Moldova). Of included studies, 10 studies focused on urban areas alone, 6 studies were located in both rural and urban areas (Cattamanchi et al., 2020; Engel et al., 2015; Joshi et al., 2018; Nalugwa et al., 2020; Shewade et al., 2018), and 5 studies did not report on urban/rural settings (Hoang et al., 2015; Mwaura et al., 2020; Newtonraj et al., 2019; Oliwa et al., 2020; Royce et al., 2014).
The included studies researched a variety of users including private and public doctors, pediatricians, nurses, community health workers, (MDR-TB) patients, household contacts, laboratory technicians, policymakers, and managers or implementers. While it is difficult to quantify the numbers of participants as not all studies reported these in detail, for those that specified number of participants there were 813 participants in total and for those that specified numbers of type of participants they involved in total 101 patients, 8 household contacts of MDR patients, 609 health care workers and TB programme managers (of which 35+ lab personnel) and 8 manufacturers.
All studies focused on Xpert/MTB RIF, except one which focused on Xpert Ultra (Mwaura, et al., 2020). The studies did not all report in detail how the diagnostic was used. Among those studies which provided details, a minority of studies reported low-complexity automated NAATs being used upfront for all presumptive TB patients (Naidoo, Colvin). In most studies, low-complexity automated NAATs were used upfront only for selected patient groups (McDowell et al., 2018; Nalugwa, et al., 2020; Newtonraj, et al., 2019; Oliwa, et al., 2020; Rendell et al., 2017; Vijayageetha et al., 2019), household contacts of MDR-TB patients (Phyo et al., 2019), for previously treated patients (McDowell & Pai, 2016; Royce, et al., 2014) and/or as a follow-up for smear negatives (Cattamanchi, et al., 2020; Creswell et al., 2014; Newtonraj, et al., 2019; Rendell, et al., 2017; Shewade, et al., 2018).
The included studies covered a range of facilities including clinics, (district) hospitals, microscopy centres, NAAT testing sites, national reference laboratories and provincial laboratories and community outreach settings. Some studies combined clinics, hospitals and NAAT testing facilities, others focused on one type of facilities alone. The majority of studies reported results from public facilities (11), five studies reported results from both public and private facilities (Creswell, et al., 2014; Engel, et al., 2015; Jaroslawski & Pai, 2012; McDowell, et al., 2018; Newtonraj, et al., 2019), one study reported results from a NGO-led project (Phyo, et al., 2019) and one from just private facilities (McDowell & Pai, 2016). Finally, two studies did not report public/private (Naidoo, van Niekerk, du Toit, Beyers, & Leon, 2015; Royce, et al., 2014).
Of included studies, seven used a mixed method design (Cattamanchi, et al., 2020; Joshi, et al., 2018; Nalugwa, et al., 2020; Phyo, et al., 2019; Royce, et al., 2014; Stime et al., 2018; Vijayageetha, et al., 2019) while the remaining 14 were purely qualitative in nature. The majority was descriptive in nature with only 6 studies explicating the use of a theoretical framework. The study objectives focused mainly on understanding the perspectives of providers, managers or patients engaged in TB diagnosis or screening

and low-complexity automated NAATs use and challenges to their implementation (Cattamanchi, et al., 2020; Creswell, et al., 2014; Jaroslawski & Pai, 2012; Joshi, et al., 2018; McDowell, et al., 2018; Mwaura, et al., 2020; Naidoo, et al., 2015; Newtonraj, et al., 2019; Phyo, et al., 2019; R. de Camargo, R. Guedes, Caetano, Menezes, & Trajman, 2015; Rendell, et al., 2017; Royce, et al., 2014; Shewade, et al., 2018). A second set of studies had a more procedural approach where understanding the process of using or implementing diagnostics was the main aim which generated data on the perspectives as well as practices of users (Colvin et al., 2015; Engel, et al., 2015; Hoang, et al., 2015; McDowell & Pai, 2016; Oliwa, et al., 2020; Stime, et al., 2018).

Methodological limitations of the studies

The sampled studies were overall of good quality with about half of them having undertaking a thorough attempt or several steps towards methodological quality across the assessed components and the other half having mostly undertaken at least a few steps towards methodological quality. Details of the assessments of methodological limitations for individual studies can be found in Appendix 2.

Confidence in the review findings

Out of 18 findings, we graded 14 as high confidence, 3 as moderate confidence and one as low confidence using the CERQual approach. For summary and explanations of our CERQual assessment see Appendix 3, CERQual Summary of Qualitative Findings table.

Review findings

From our synthesis, we developed 18 individual findings, which we organised into three overarching categories related to 1) critical aspects users value, 2) challenges to realizing those values and 3) concerns for access and equity. In the sections below we present each finding followed by the detailed results. We developed a figure to illustrate how these findings interact (see conceptual model and Figure 2).

Critical aspects users value

Summary of Qualitative Findings table: finding 1-6 (see appendix 3)

Finding 1: Patients in high TB burden countries value 1) getting an accurate diagnosis and reaching diagnostic closure (finally knowing what is wrong with me), 2) avoiding diagnostic delays as they exacerbate existing financial hardships and emotional and physical suffering and make patients feel guilty for infecting others (especially children), 3) having accessible facilities and 4) reducing diagnosis-associated costs (travel, missing work) as important outcomes of the diagnostic. (moderate confidence (Joshi, et al., 2018; Naidoo, et al., 2015; Phyo, et al., 2019; R. de Camargo, et al., 2015; Royce, et al., 2014; Vijayageetha, et al., 2019). Even though a MDR-TB diagnosis is devastating for patients, patients value reaching diagnostic closure through an accurate diagnosis and finally knowing what is wrong with them (Naidoo, et al., 2015). MDR-TB patients highlighted how diagnostic delays exacerbate existing financial and other hardships or create new ones (avoidable delays that lead to emotional and physical suffering and onwards transmission of MDR-TB to children). Diagnostic delays make patients feel guilty of infecting others and they experience distress when on first line treatment that does not help (Naidoo, et al., 2015).

> An MDR-TB patient in South Africa highlights this: *" it hurts me a lot, I don't even want to go there, I am feeling very bad, very, very bad, because if this was detected earlier I was not going to go through some difficulties that I went through. You know… when I think that I even infected my child it makes me feel very bad. Because if this was detected early and [I was] started on the right treatment, maybe some of the problems would have been eliminated" (LPA-3)(Naidoo, et al., 2015).*

Reducing time to diagnosis and saving cost (including travel cost) is important for patients (Joshi, et al., 2018; Phyo, et al., 2019; Royce, et al., 2014; Vijayageetha, et al., 2019). In South Africa, for instance,

patients would recommend family and friends to avoid private sector services and instead immediately go to the public primary care facilities despite perceptions of long waiting times, lack of privacy and poor staff attitudes associated with public facilities (Naidoo 2015). In a study from Brazil, patients did not struggle with delays or cost, because they either lived close by testing facilities and those who did have to take a long bus trip were on medical leave (i.e. had time) and got the bus ticket subsidized (no extra cost) (R. de Camargo, et al., 2015).

Finding 2: Compared to existing tests/sputum microscopy, healthcare professionals appreciate the rapidity and accuracy of low-complexity automated NAAT results, the diversity of sample types, ability to detect drug resistance, as well as the consequence of avoiding costlier investigations or hospital stays when using low-complexity automated NAATs. (high confidence: (Joshi, et al., 2018; McDowell, et al., 2018; Mwaura, et al., 2020; Naidoo, et al., 2015; Newtonraj, et al., 2019; R. de Camargo, et al., 2015; Rendell, et al., 2017; Vijayageetha, et al., 2019). Several studies mentioned how healthcare workers value the time saving potential of low-complexity automated NAATs when receiving results more quickly (Joshi, et al., 2018; Mwaura, et al., 2020; Newtonraj, et al., 2019; Rendell, et al., 2017). Especially if same-day results allow same day treatment initiation, this is considered a vast improvement (McDowell, et al., 2018). What is more, healthcare professionals valued the ability for diagnosis in pauci-bacillary samples (Joshi, et al., 2018; Newtonraj, et al., 2019) and in a diversity of sample types, especially important for diagnosis of children (McDowell, et al., 2018). Also, the accuracy and reliability of results is considered an important benefit (R. de Camargo, et al., 2015; Vijayageetha, et al., 2019), and, with it, particularly for Xpert Ultra, the improved TB case detection among hard-to-diagnose patients (Mwaura, et al., 2020), less ordering of other expensive investigations (CT scan, bronchoscopies) and avoidance of longer hospital stays for children (McDowell, et al., 2018).
According to one study, it is the experience of using low-complexity automated NAAT and of its added value (especially speed, affordability and generation of additional insights or increased confidence in results) that drives behaviour change among clinicians, more than education and information about the product(McDowell, et al., 2018).

Finding 3: Low-complexity automated NAAT allows healthcare workers to detect drug resistance earlier and pediatricians in particular mentioned how it heightened their risk perception of drug resistance in children; yet in a context with widespread severe forms of drug resistance and a habit of treating empirically first, clinicians see the inability to detect resistance of some NAATs beyond rifampicin as a hindrance. (high confidence (Joshi, et al., 2018; McDowell & Pai, 2016; McDowell, et al., 2018; Naidoo, et al., 2015; R. de Camargo, et al., 2015). The ability to detect drug resistant TB early is appreciated among healthcare workers (Joshi, et al., 2018; R. de Camargo, et al., 2015). Particularly in children, where physicians do not typically expect drug resistance, low-complexity automated NAATs' use altered physicians' risk perception of MDR-TB in children and reduced empirical treatment among children (McDowell, et al., 2018). Yet, in a context of severe forms of drug-resistant TB and where treating empirically is common, such as in the private sector in Mumbai, India, the added value of low-complexity automated NAATs that only detect rifampicin resistance is questioned.

> "*An MBBS doctor in Mumbai commented: But why should I use Xpert? It only tells if the patient is rifampicin susceptible or not, but it does not tell me anything else. It is better to give first-line drugs and see if the patient responds. After some time we will know if the first-line drugs are working and if they do not we know we need to move on. Xpert tells us about rifampicin quickly but what we really need is a culture and that takes time. In Mumbai, Xpert is not enough to decide on a proper second-line regimen." (McDowell & Pai, 2016).*

Finding 4: Clinicians value the confidence that low-complexity automated NAAT results generate, to start treatment, to reassure and motivate patients and their caretakers, to justify actions towards other doctors and to increase collaboration between private/public providers. (high confidence (McDowell, et al., 2018; Oliwa, et al., 2020). Having confidence in diagnostic results is valued as important

to start treatment, reassure and motivate family and patients (to begin and adhere to treatment) and justify actions towards other doctors who do not question low-complexity automated NAAT result. Experience with successful treatment following a positive NAAT result increases that confidence among clinicians and pediatricians (McDowell, et al., 2018; Oliwa, et al., 2020). Availability of low-complexity automated NAAT and fast turnaround times also increased confidence in overall quality of public sector laboratories and willingness to collaborate and refer patients among private pediatricians in India (McDowell, et al., 2018).

Finding 5: Laboratory technicians appreciate the improvement of overall laboratory work that low-complexity automated NAAT brings compared to sputum microscopy in terms of ease of use, ergonomics, and biosafety. (high confidence:(Creswell, et al., 2014; Newtonraj, et al., 2019; R. de Camargo, et al., 2015). The improved laboratory conditions work as an incentive for workers (Creswell, et al., 2014; R. de Camargo, et al., 2015).

Finding 6: Laboratory managers appreciate that monitoring of laboratory work and training is easier than with sputum microscopy and that low-complexity automated NAAT eases staff retention, as it increases staff satisfaction and has a symbolic meaning of progress within the TB world. (low confidence:(R. de Camargo, et al., 2015). According to laboratory managers in Brazil, low-complexity automated NAAT has a symbolic meaning in the tuberculosis diagnostic field that has spent decades without innovation: "*The emotional and psychological factors of the workers who will be most pleased to do its work, will get sick less often, take fewer licenses, will be less prone to giving up working in that area. We saw a great satisfaction*." (Manager 1, Manaus) (R. de Camargo, et al., 2015).

Challenges to realizing those values

Summary of Qualitative Findings table: finding 7-15 (see appendix 3)

Finding 7: Patients can be reluctant to test for TB/MDR-TB because of stigma related to MDR-TB or related to having interrupted treatment in the past, because of fears of side effects, the failure to recognize symptoms, the inability to produce sputum and the cost, distance and travel concerns related to (repeat) clinic visits. (high confidence: (Naidoo, et al., 2015; Phyo, et al., 2019; Royce, et al., 2014; Shewade, et al., 2018).

Associated stigma, discriminatory attitudes at clinics and fear prevent patients from returning for providing second sputum for DST according to healthcare workers in India (Shewade, et al., 2018). The fear of treatment associated side effects can prevent patients from testing (Phyo, et al., 2019)Stigma can lead to misclassifications and delayed DST. In a study on healthcare workers perspectives on potential barriers to the detection of MDR-TB in previously treated TB patients in Cambodia, healthcare workers mentioned how patients are ashamed to reveal previous interrupted treatment to health workers leading to misclassification (Royce, et al., 2014). *Some participants noted that .. . many patients hide their previous treatments . .. they are ashamed [of revealing that they interrupted treatment previously] ((Royce, et al., 2014) p.1303).* Healthcare workers also observed how patients are afraid to reveal a MDR-TB diagnosis to others at home (Royce, et al., 2014).

Failure to recognize symptoms (not as TB related, or associating them with HIV instead) and denying or minimizing symptoms can lead to delays and explain why many patients are very ill at first contact (Naidoo, et al., 2015).

The inability to produce sputum and not having symptoms can prevent contacts from MDR-TB patients to agree being tested (Phyo, et al., 2019). Inability to produce sputum after a certain period of TB treatment could be a reason why patient did not return with two specimens for DST (esp because of the delays between initial TB diagnosis and DST) (Shewade, et al., 2018).

Long distances, financial constraints and inconvenient clinic hours can prevent patients from testing (Phyo, et al., 2019; Royce, et al., 2014).

Finding 8: Health workers can be reluctant to test for TB or MDR-TB because of TB associated stigma and consequences for their patients, fears of acquiring TB, fear from supervisors when reclassifying patients already on TB treatment who turn out to be misclassified, fear of side effects of drugs in children, and community awareness of disease manifestations in children. (high confidence:(Oliwa, et al., 2020; Royce, et al., 2014).

In the context of pediatric TB in Kenya, health workers can be reluctant to test for TB because of the association of TB with being HIV positive. This makes healthcare workers worry about the emotional burden a diagnosis would inflict upon their patients, which the following quote of a pediatrician illustrates:

> *" … And then there is that thing people thinking TB is equal to HIV, so when now someone has been told that they have TB now everyone thinks that they are HIV positive, so there is that even being shunned by the family. I have a mother right now who was actually chased away by her extended family because of the TB diagnosis…" Paediatrician_SSI_03 ((Oliwa, et al., 2020), p8)*

Additionally, the fear of acquiring TB as health worker, the fear of side effects of drugs in children, and the belief that children do not get TB contribute to underutilization of TB diagnostics (Oliwa, et al., 2020). In another study, health workers did not want to test for DST because it would mean to reclassify patients (if it emerged that they had been previously treated) because of fear of what supervisor would think when controlling the register and discovering a change. If previously treated patients are not registered accordingly, it delays DST (Royce, et al., 2014).

Finding 9: Diagnostic delays are accumulated because of various health system factors (i.e. non-adherence to testing algorithms, testing for (MDR-)TB late in the process, empirical treatment, false negatives due to technology failure, large sample volumes and staff shortages, poor/delayed sample transport and result communication, delays in scheduling follow up visits and recalling patients, inconsistent result recording) and to a lesser extent patient-related delays (i.e. missed follow-up appointments, competing family demands and seeking traditional health-care). (high confidence: (Cattamanchi, et al., 2020; Creswell, et al., 2014; Engel, et al., 2015; McDowell & Pai, 2016; Naidoo, et al., 2015; Nalugwa, et al., 2020; Rendell, et al., 2017; Royce, et al., 2014; Stime, et al., 2018).

Rapid turn-around time is an important potential of diagnostic algorithms involving automated NAATs of low complexity and an important outcome for health providers and patients for some providers determining utilization of these diagnostics. Users value receiving results more quickly to speed up clinical work and to free time in the laboratory while a cycle is running (R. de Camargo, et al., 2015). The potential of an algorithm involving automated NAATs to reduce diagnostic delays is emphasized across studies (Naidoo, et al., 2015; Newtonraj, et al., 2019; Rendell, et al., 2017) and illustrated with two examples in Naidoo 2015 where rapid initiation of MDR-TB treatment happened, within 6 and 8 days of the first health contact, respectively. "*Early access to treatment was enabled by the correct tests being requested which yielded a positive result, results being available when patients returned and decentralised treatment being available*." (Naidoo, et al., 2015)

But in many places the overall turn-around time of low-complexity automated NAATs is increased due to accumulation of delays and how diagnostic and treatment algorithms are organized. Many authors differentiate between health system factors and patient-related delays:

Health system factors include: failure to adhere to testing algorithms (providers not testing for TB or MDR-TB at initial visits; correct tests not initially done (Naidoo, et al., 2015) or providers preferring empirical treatment over testing (McDowell & Pai, 2016), failure in the testing technology [false negatives mostly], problems with receiving the results, scheduling follow-up visits and recalling patients with positive results (Cattamanchi, et al., 2020; Naidoo, et al., 2015; Nalugwa, et al., 2020); increased turn-around times due to large number of samples being tested and machines not running over night (Engel, et al., 2015; Stime, et al., 2018); staff shortages (Stime, et al., 2018), delays in transporting samples to NAAT testing sites and in reporting/receiving results back (limited communication possibilities via phone, sms, overextended courier system, or reliance on paper-based system), inconsistent recording of TB results at facilities (Creswell, et al., 2014; Nalugwa, et al., 2020; Royce, et al., 2014); lack of follow-up system when patients are being referred to testing sites (Engel, et al., 2015). In Moldova participants reported a delay (1-2 weeks) in

initiating MDR-TB treatment because of the procedural requirement to determine the MDR-TB treatment plan at a weekly consensus meeting (Rendell, et al., 2017).
In South Africa, "*delays overall were longer for patients in whom initial tests were negative with 1st-line TB treatment started on clinical or chest x-ray findings*."(p9) (Naidoo, et al., 2015). Strategies by providers to deal with associated delays create new problems such as artificially prolonging turnaround times when asking patients to come back later, anticipating delays (Engel, et al., 2015). Passage of time and multiple failed empirical broad-spectrum antibiotic trials are necessary before private practitioners in India consider TB, resulting in **l**ong delays in diagnosing TB (McDowell & Pai, 2016).
Patient-related delays contributed to a lesser extent, but can happen due to not recognizing symptoms, missed follow-up appointments, competing family demands and seeking traditional health-care)." (Naidoo, et al., 2015).

Finding 10: Poor sample quality, inconvenient sample collection facilities, non-functioning sample transport mechanisms, and difficulty of obtaining pediatric samples can cause error results and underutilization of low-complexity automated NAAT. (high confidence: (Cattamanchi, et al., 2020; Creswell, et al., 2014; Hoang, et al., 2015; McDowell & Pai, 2016; McDowell, et al., 2018; Newtonraj, et al., 2019; Oliwa, et al., 2020; Phyo, et al., 2019; Rendell, et al., 2017; Royce, et al., 2014; Shewade, et al., 2018; Vijayageetha, et al., 2019).
Providers struggle with poor sample quality causing error results (Newtonraj, et al., 2019; Rendell, et al., 2017; Royce, et al., 2014). Reasons can be delays and inadequate specimen transportation, insufficient instructions for patients (Creswell, et al., 2014)and collecting sputum many days after retreatment initiation (by which time the cough may be resolved and it is harder to provide a specimen (McDowell & Pai, 2016; Royce, et al., 2014) and specimen have very low bacteria count (McDowell & Pai, 2016). Convenient sample collection facilities and functioning sample transport are essential to ensure utilization of Xpert and avoid delays (Cattamanchi, et al., 2020; Creswell, et al., 2014; Hoang, et al., 2015; Newtonraj, et al., 2019; Phyo, et al., 2019; Shewade, et al., 2018; Vijayageetha, et al., 2019). Among the involved studies, turn-around times of low-complexity automated NAATs range from same day to 1-2 weeks. In India, healthcare workers reported difficulties in convincing patients to produce two sputum samples for low-complexity automated NAAT if sputum was negative or they had to travel long distances to come for xray and then might not be able to return again for second sample. Sample collection facilities would be more convenient if patients could provide sputum specimens at the nearest primary healthcare clinic (Newtonraj, et al., 2019). At a public MDR treatment programme in Vietnam the lack of a functioning sputum transport system (no appropriate financial compensation mechanisms for consumable procurement and transportation fees; no agreements with postal services, using health staff and public transport instead) led to underutilization of NAAT machines (Hoang, et al., 2015). In India, the lack of an assured specimen transport after patient identification required coordinating health worker (to transport sample) and returning patients (to provide samples) to be present on same day which was challenging (Shewade, et al., 2018).

Finding 11: The lack of sufficient resources and of ensuring maintenance (i.e. stock-outs; unreliable logistics; lack of funding, electricity, space, air conditioners, and sputum containers; dusty environment, and delayed or absent local repair option) leads to higher test failure rates and underutilization and negatively influences uptake and impact of low-complexity automated NAAT. (high confidence:(Creswell, et al., 2014; Hoang, et al., 2015; Joshi, et al., 2018; Mwaura, et al., 2020; Nalugwa, et al., 2020; Oliwa, et al., 2020; R. de Camargo, et al., 2015; Rendell, et al., 2017; Shewade, et al., 2018; Stime, et al., 2018)
For instance, several studies reported stock outs and unreliable logistics around cartridges among common resource challenges at sites running Xpert (Hoang, et al., 2015; Joshi, et al., 2018; Mwaura, et al., 2020; Nalugwa, et al., 2020; Oliwa, et al., 2020; Rendell, et al., 2017). "*Stock-outs (..) led to delays in making a diagnosis and reinforced a reluctance in ordering the tests in future. This shows how age-old system issues like stock-outs potentially affect adoption of new diagnostics*." (Oliwa, et al., 2020).

Poor laboratory infrastructure, including frequent power cuts, lack of air conditioners and/or dusty environment and lack of adequate rooms or proper furniture, can challenge proper testing (Creswell, et al., 2014; Joshi, et al., 2018; Nalugwa, et al., 2020) and explain high test failure rates and indeterminate results (Joshi, et al., 2018). Yet, differences between types of failed tests are unclear and available data not always used. 'No result' test results were often caused by a power failure (Creswell, et al., 2014).
There was need for more basic office equipment including functioning internet connections to cater for the introduction of new equipment (R. de Camargo, et al., 2015; Rendell, et al., 2017; Shewade, et al., 2018). Sputum collection facilities in a hub and spoke model struggled with lack of sputum transport containers and lack of electricity to enable refrigeration. This meant patients needed to come back to provide a sputum sample on transport day, which many patients would not do (Nalugwa, et al., 2020).
Delays in calibration and replacement of damaged modules (Creswell, et al., 2014; Joshi, et al., 2018) and absence of local repair options challenge sustainability of low-complexity automated NAAT. A study from Mongolia, for instance, reported difficulties for arranging repairs when required because of limited availability of trained mechanics and how having internal capacity for repair helps to prevent interruption of workflows (Rendell, et al., 2017).

Finding 12: Low-complexity automated NAAT seems to decrease workload by freeing up time for laboratory staff, but in most settings staff may be hesitant to accept testing with low-complexity automated NAAT because it increases workload if added onto existing laboratory work without adjusting staffing arrangements, or if it does not replace existing diagnostic tests. (moderate confidence: (Joshi, et al., 2018; Phyo, et al., 2019; R. de Camargo, et al., 2015; Rendell, et al., 2017; Shewade, et al., 2018; Stime, et al., 2018; Vijayageetha, et al., 2019).
In settings where low-complexity automated NAAT is introduced without replacing existing diagnostics or adequate staffing arrangements, it generates more work for laboratory technicians (Joshi, et al., 2018; R. de Camargo, et al., 2015; Rendell, et al., 2017): "Because you're working with two methods instead of one." (R. de Camargo, et al., 2015). This can then mean that contact investigations of MDR-TB patients are not done (Phyo, et al., 2019), there is a lack of accountability in tracking of patients after identification and referral (Shewade, et al., 2018) and that staff is hesitant to accept POC testing with Xpert (Stime, et al., 2018). Lack of dedicated staff and high workload of existing staff is also hindering implementation of screening for TB among pregnant women (Vijayageetha, et al., 2019).

Finding 13: Workflows, professional roles and patient flows matter for utilizing low-complexity automated NAAT, for instance inefficient organizational processes, poor links between providers, unclear follow up mechanisms or where patients need to go can delay diagnostic processes. (high confidence: (Hoang, et al., 2015; Oliwa, et al., 2020; R. de Camargo, et al., 2015; Royce, et al., 2014; Stime, et al., 2018)
The introduction of low-complexity automated NAATs often has implications for workflows and professional roles. These matter for acceptance by the users. In Brazil, the introduction of low-complexity automated NAAT brought a change in workflow where the lab technician, after examining the quality of the sputum sample, decides if the sample can be tested on low-complexity automated NAAT or sputum microscopy (samples with low volume and samples with food/blood residues cannot go on low-complexity automated NAAT). This change in workflow did not translate into a change in professional roles, the lab technician remained responsible for the entire process including authorizing the delivery of results. The authors argued that this meant the lab technicians more easily accepted the technology (R. de Camargo, et al., 2015). Existing inefficient workflows can also cause delay in making automated NAATs of low complexity work, for instance when reporting results through paper-based (Rendell, et al., 2017; Royce, et al., 2014), or non-standardized systems without clear guidance or accountability (Creswell, et al., 2014; Shewade, et al., 2018), incorrect filing of patient documents (Stime, et al., 2018), unclear follow up mechanisms (Oliwa, et al., 2020; Stime, et al., 2018), poor links between public and private providers (Hoang, et al., 2015) or when laboratory technicians have limited hours available to work (Creswell, et al., 2014).

Low-complexity automated NAAT use also has implications for patient flows who only have to submit one sputum sample (Phyo, et al., 2019) but might find it difficult to find their way through different sites and departments (Stime, et al., 2018; Vijayageetha, et al., 2019) or where they need to (Oliwa, et al., 2020; Stime, et al., 2018).

Finding 14: Too much confidence in low-complexity automated NAAT's accuracy can mean blindly accepting results without using clinical impressions or for patients to trust low-complexity automated NAAT because it is a computer-based result (moderate confidence: (Joshi, et al., 2018; Mwaura, et al., 2020; Newtonraj, et al., 2019).

Owing to the confidence in low-complexity automated NAAT's accuracy clinicians accept negative results without using clinical impressions to question these and are missing patients (Mwaura, et al., 2020; Newtonraj, et al., 2019). Xpert is taken as a gold standard and TB is ruled out, without being aware that results may vary in extra-pulmonary TB or poor quality samples and that there might be false negatives (Newtonraj, et al., 2019). Clinicians in Kenya and Eswatini anticipated that with Xpert MTB/RIF Ultra this tendency would increase, empirical diagnosis would further decrease while the number of bacteriological confirmed cases would increase among these hard-to-diagnose patient groups because of the trace calls (Mwaura, et al., 2020). One study reported that a computer-based test generates confidence in patients:

> "*Patients also prefer Xpert test thinking it will give an accurate result because it is computer-based. They will go for test (i.e. Gene X-pert). Patients demand to test by machine/ computer. They have trust towards Gene X-pert. Even though we only test by X-pert if referred by physician.'* (X-pert staff) (Joshi, et al., 2018)

Finding 15: Implementation processes have been challenged by lack of data on pragmatic effectiveness in operational conditions, lack of knowledge and awareness among providers beyond lab personnel, lack of guidelines and standardized training modules and instructions and a lack of national policy consensus and inclusive decision-making prior to roll out. (High confidence: (Colvin, et al., 2015; Creswell, et al., 2014; Hoang, et al., 2015; Joshi, et al., 2018; Naidoo, et al., 2015; Newtonraj, et al., 2019; R. de Camargo, et al., 2015; Rendell, et al., 2017; Shewade, et al., 2018).

Generating data on how new diagnostics should best and most effectively be integrated into local operational context of use including practical feasibility planning is crucial prior to implementation as well as during early implementation to inform roll out and impact on TB control (Colvin, et al., 2015; Joshi, et al., 2018). The early Xpert MTB/RIF (and LPA) demonstration studies in South Africa were assessing accuracy but not pragmatic effectiveness in operational conditions which is a missed opportunity (Colvin, et al., 2015).

When introducing new diagnostics, several studies cited challenges with ensuring knowledge and awareness about the diagnostic and guidelines (Colvin, et al., 2015; Joshi, et al., 2018; Newtonraj, et al., 2019; Rendell, et al., 2017; Shewade, et al., 2018) not only among laboratory technicians or managers, but also among the public, clinicians and health workers (Colvin, et al., 2015). Lack of clear and updated guidelines and poor dissemination at lower levels and among private providers challenges implementation (Creswell, et al., 2014; Hoang, et al., 2015; Newtonraj, et al., 2019; Rendell, et al., 2017). Clinicians should be included in trainings (R. de Camargo, et al., 2015). If not done this led to poor referral to low-complexity automated NAAT (Joshi, et al., 2018; Newtonraj, et al., 2019; Shewade, et al., 2018) or inconsistency in what samples were used (Rendell, et al., 2017). In the TB REACH projects, for instance, staff rotation and new practices around request forms, specimen transport and clinical decisions for rifampicin resistant results posed crucial training challenges (Creswell, et al., 2014). In Vietnam, a lack of standardized training modules and instructions led to failures in identifying presumptive patients, especially among risk groups. Training and skill development was further challenges by high staff turnover or changes in staff responsibilities (Hoang, et al., 2015). Insufficient attention to change management processes at facility level can hamper impact of diagnostic (Colvin, et al., 2015; Naidoo, et al., 2015). In South Africa, changes to TB testing algorithms, laboratory request forms, and national TB registers happened only later after implementation (Colvin, et al., 2015).

When introducing new diagnostics, it is also important to include relevant stakeholders in decision-making processes and in planning regarding implementation and allow a national policy consensus process. This could involve national and provincial programme managers and health officers, clinicians, and laboratory staff. In South Africa, Xpert had high visibility but its introduction was not inclusive, focused around FIND, WHO, NHLS, and the ministry of health, sidelining key national and provincial actors in the TB programme. The lack of inclusion and communication was perpetuated by the fast pace of implementation and high international pressure to act (rescue vs management). (Colvin, et al., 2015)

> *"TB managers and local health services staff alike experienced the decision making about and implementation of Xpert as fast-paced, with little horizontal co-ordination or communication, although Xpert involved more on-the-ground changes than LPA. (…) The rapid pace of implementation meant there was little time to assess its operation and integration into local contexts, and in the words of one manager, many staff felt that Xpert seemed to have just 'fallen out of the sky' at a time when their focus was still on the completion of the LPA rollout." ((Colvin, et al., 2015) p.1333)*

Concerns for access and equity

Summary of Qualitative Findings table: finding 16-18 (see appendix 3)

Finding 16: Staff and managers voiced concerns regarding sustainability of funding and maintenance, complex conflicts of interest between donors and implementers and concerns related to the strategic and equitable use of resources, which negatively affects creating equitable access to automated NAATs of low complexity. (High confidence: (Colvin, et al., 2015; Creswell, et al., 2014; Jaroslawski & Pai, 2012; R. de Camargo, et al., 2015).

Staff and managers expressed concerns about high cost and sustainability of low-complexity automated NAATs (Colvin, et al., 2015; Creswell, et al., 2014; R. de Camargo, et al., 2015) and the challenges of funding maintenance of the devices (R. de Camargo, et al., 2015). Donor funding might have led to insufficient attention being paid to ongoing resource requirements (i.e. masking startup and recurrent cost, appearing more feasible) (Colvin, et al., 2015). Affordability is crucial for utilization of diagnostics in the private sector in India, where prices are often inflated, which meant that inadequate alternatives such as serology were preferred by patients, laboratory technicians and clinicians over molecular tests (Jaroslawski & Pai, 2012).

Participants in a study in South Africa voiced concerns about strategic and equitable use of resources, because low-complexity automated NAAT was placed in hospitals (which already have LPA) and selected, often well-functioning sub-districts and not in primary health clinics or areas with no access to improved TB diagnostics. The decision of where to deploy automated NAATs of low complexity, was not made by provincial and district managers (Colvin, et al., 2015). Complex conflict of interest created dependence on a single provider for crucial health technologies. Colvin and colleagues recommend to carefully manage these conflicts prior and during development and implementation of diagnostics:

> *"Among our recommendations is the need to identify and manage conflicts of interest that may arise when innovative partnerships are established to address public health issues. We suggest that the role of committed leadership in fast-tracking processes needs to be matched with a national policy consensus process and careful, transparent planning."((Colvin, et al., 2015)p.1337)*

Finding 17: Lengthy diagnostic delays, underutilization of automated NAATs of low complexity, lack of TB diagnostic facilities at lower levels and too many eligibility restrictions, hamper access to prompt and accurate testing and treatment particularly for vulnerable groups. (High confidence: (Engel, et al., 2015; Hoang, et al., 2015; Joshi, et al., 2018; McDowell & Pai, 2016; McDowell, et al., 2018; Naidoo, et al., 2015; Nalugwa, et al., 2020; Newtonraj, et al., 2019; Oliwa, et al., 2020; Phyo, et al., 2019; Royce, et al., 2014)

Several studies showed how lengthy diagnostic delays, underutilization of low-complexity automated NAAT and lack of TB diagnostic facilities at lower levels where many presumptive patients present, hamper access to prompt and accurate treatment for those that are eligible for testing (Nalugwa, et al., 2020) with vulnerable groups and patients with difficult disease patterns (including children (Joshi, et al., 2018; McDowell, et al., 2018; Oliwa, et al., 2020), MDR-TB (Hoang, et al., 2015; Naidoo, et al., 2015)) or patients with limited ability to pay (for fees or transport cost to overcome distance and produce second sample) (Engel, et al., 2015; Joshi, et al., 2018; McDowell & Pai, 2016; Newtonraj, et al., 2019; Phyo, et al., 2019; Royce, et al., 2014)) affected the worst. Limited ability to pay means private providers treat rather than order tests (McDowell & Pai, 2016). Deployment and eligibility decisions and overcoming challenge to diagnostic delay and underutilization are crucial in enabling access.

> *"Only when each pediatric presumptive TB patient is offered upfront Xpert testing a more synchronized pediatric TB case management, same day TB diagnosis, and access to prompt and accurate TB treatment can be guaranteed. Locating Xpert at the end in the diagnostic process or placing too many restrictions on the criteria of patients who can access the test will limit its impact significantly"*. ((McDowell, et al., 2018), p13).

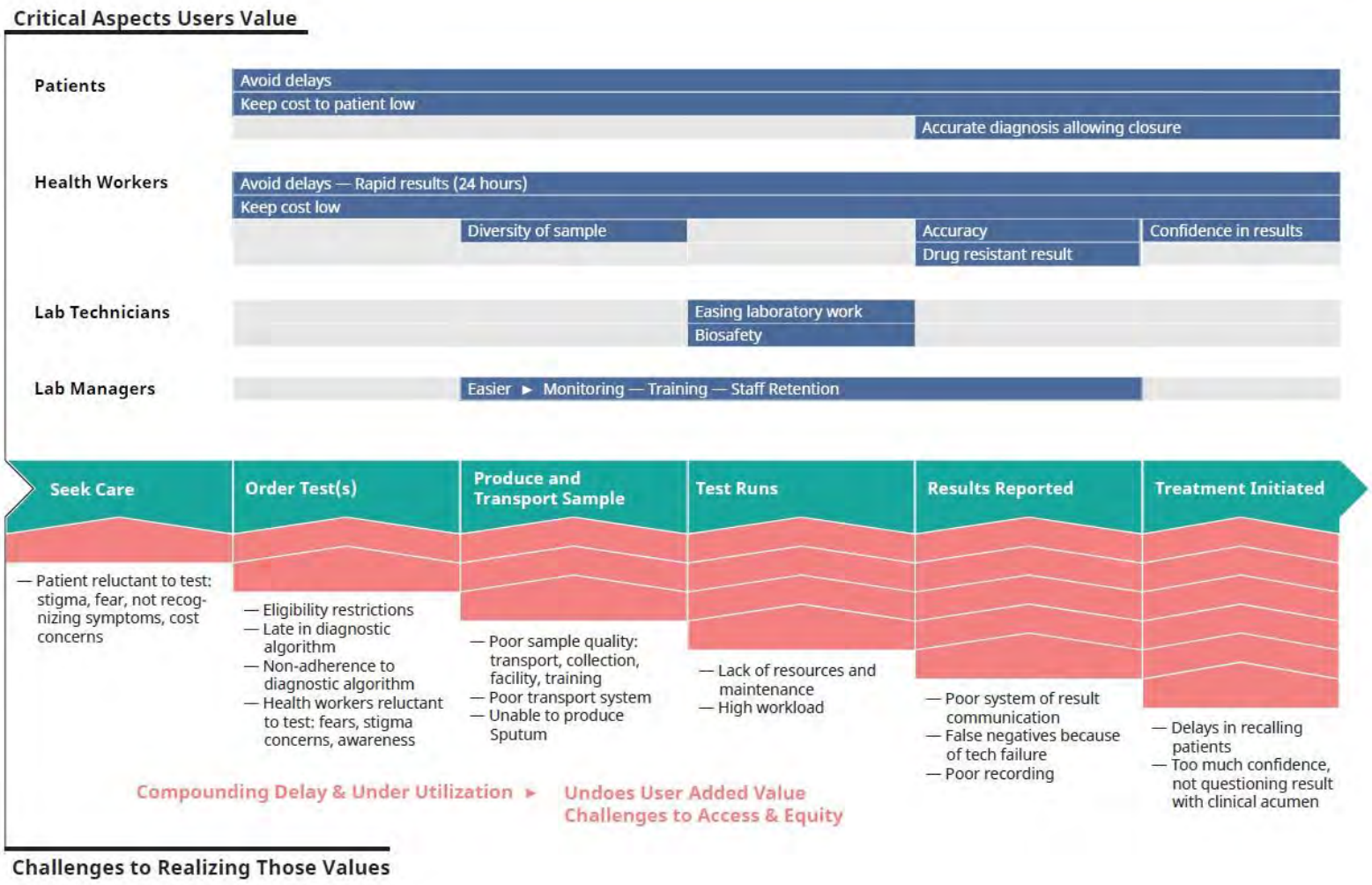
Finding 18: The identified challenges and accumulated delays risk undoing the added value as identified by the users, ultimately leading to underutilization and important implications for access and equity. (High confidence: review finding #1-15, (Engel, et al., 2015; McDowell, et al., 2018; Naidoo, et al., 2015; Shewade, et al., 2018)

The challenges identified in findings 7-15 risk undoing the added value as identified by different users in findings 1-6. For instance, diagnostic delays can further compound underutilization of low-complexity automated NAAT and risk patient loss from diagnostic and treatment pathways. An overall turnaround time within 24hrs (including transportation mechanisms and quick reporting of results electronically) was essential for use of low-complexity automated NAAT among pediatricians. The impact of low-complexity automated NAAT with longer TAT is less certain (McDowell, et al., 2018). The delays between initial TB diagnosis and DST mean that some patients are unable to produce sputum after a certain period of first line TB treatment and therefore will not return with the second specimen for DST (Shewade, et al., 2018). Patient and health system delays interact. Professionals responding to anticipated health system delay, create further delays to avoid additional patient delays and patients having to wait or return again later if results are not yet available (Engel, et al., 2015).

Conceptual Model:

Based on these review findings, we have summarized how these findings interact in a conceptual model illustrated in Figure 2. The upper half of the figure illustrates the critical aspects that patients, healthcare workers, laboratory technicians and managers value (review findings 1-6). These aspects are mapped along a simplified fictive process of using low-complexity automated NAAT, consisting of the following steps: seek care, order test(s), product and transport sample, test runs, results reported and treatment initiated. The length of the blue bars indicates at what step in the process these user values matter (it does *not* indicate a weighted importance). The lower half of the figure illustrates the challenges to realizing those values that we identified (review findings 7-15). These challenges are listed per step in the diagnostic process at which they happen, meaning some review findings cover several steps (i.e. review finding 9 on diagnostic delays). At every step, these challenges compound diagnostic delay and underutilization of low-complexity automated NAAT with important implications for access and equity (review findings 17, 18). And at every step, these challenges risk undoing the added values that users perceive low-complexity automated NAATs offer (review finding 18). We can assume that if these values are not met users are less likely to find low-complexity automated NAATs acceptable.

Figure 1


Critical Aspects Users Value
Patients
Avoid delays
Keep cost to patient low
Accurate diagnosis allowing closure
Health Workers
Avoid delays — Rapid results (24 hours)
Keep cost low
Diversity of sample
Accuracy
Drug resistant result
Confidence in results
Lab Technicians
Easing laboratory work
Biosafety
Lab Managers
Easier ► Monitoring — Training — Staff Retention
Seek Care
Order Test(s)
Produce and Transport Sample
Test Runs
Results Reported
Treatment Initiated
— Patient reluctant to test: stigma, fear, not recognizing symptoms, cost concerns
— Eligibility restrictions
— Late in diagnostic algorithm
— Non-adherence to diagnostic algorithm
— Health workers reluctant to test: fears, stigma concerns, awareness
— Poor sample quality: transport, collection, facility, training
— Poor transport system
— Unable to produce Sputum
— Lack of resources and maintenance
— High workload
— Poor system of result communication
— False negatives because of tech failure
— Poor recording
— Delays in recalling patients
— Too much confidence, not questioning result with clinical acumen
Compounding Delay & Under Utilization ► Undoes User Added Value
Challenges to Access & Equity
Challenges to Realizing Those Values

Discussion and conclusion

This review synthesized qualitative research on user perspectives and experiences with automated NAATs of low complexity for detection of tuberculosis and tuberculosis drug resistance. We organized the 18 individual review findings into the following three overarching categories:

1) critical aspects users value: Patients value reaching diagnostic closure with an accurate diagnosis, avoiding delays, accessible facilities, keeping cost low. Healthcare professionals similarly value aspects of accuracy the resulting confidence in TB-NAAT results, rapid turnaround times and keeping cost low, as well as diversity of sample types and drug resistance information. Laboratory personnel appreciated the improvement of laboratory work and increased staff satisfaction.

2) challenges to realizing those values included patients and healthcare workers being reluctant to test for (MDR)TB due to fears, stigma, or cost concerns; accumulation of diagnostic delays at every step due to mainly health system factors; poor sample quality/transport; lack of sufficient resources and maintenance, increased workload, inefficient/unclear work- and patient flows, overreliance on low-complexity automated NAAT results at the expense of clinical acumen; and lack of data-driven and inclusive implementation processes. These challenges lead to delays and/or underutilization.

3) Concerns for access and equity included concerns over sustainability of funding and maintenance of low-complexity automated NAAT, conflicts of interest and equitable use of resources minimizing access to low-complexity automated NAATs. Also lengthy diagnostic delays, underutilization of low-complexity automated NAAT, lack of TB diagnostic facilities at lower levels and too many eligibility restrictions, hamper access to prompt and accurate testing and treatment particularly for vulnerable groups. Overall, the identified challenges risk undoing the added values as identified by users and compounding underutilization.

Furthermore, the review finds that use of low-complexity automated NAAT is diverse but rarely used upfront for all presumptive TB patients. WHO's policy of Xpert for all patients is insufficiently implemented (England, Masini, & Fajardo, 2019).

All sampled studies included in this review were conducted in low- and middle-income countries but only one study involved a country in Eastern Europe. More research from that region could have added additional insights given the high burden of MDR-TB in the region.

The majority of the studies used interview or focus group methods while only three also used observations. It may be useful to make more use of longer-term ethnographic methods, such as observations, to better understand processes and practices of using low-complexity automated NAATs.

The multidisciplinary author team brought a substantive, contextual and methodological expertise to this review. Our findings were strengthened by a detailed, rigorous and iterative process of data-extraction and analysis involving the entire author team and a considerable body of evidence presented in this synthesis. We included studies from across different high-burden TB countries and did not identify any major themes that only occurred in one specific setting, making these findings generalizable to countries with high TB burden and low-complexity automated NAAT testing. Because we sampled a subset of the richest studies, the review remains incomplete.

Overall, the review results mean that the promise to overcome absent laboratory infrastructure and skilled human resources with point-of-care diagnostics is misleading and in fact undermines the added value new diagnostics of low complexity can bring for TB patients and healthcare professionals. Testing in more peripheral settings still requires strong health systems, laboratory infrastructure and human resources, albeit in slightly different forms (Beisel, Umlauf, Hutchinson, & Chandler, 2016; Engel et al., 2016). This means infrastructure strengthening and innovation of affordable and available diagnostic technologies needs to happen jointly (Kelly-Cirino et al., 2019).

This review also underlines earlier calls for the importance of improving implementation processes of new diagnostics (Albert et al., 2016), including early and inclusive engagement of in-country stakeholders from different levels, broader systems strengthening, improved data on ground level realities prior and during implementation, as well as pro-active management of conflicts of interests in order to ensure equitable use

of resources. Implementation processes that do not pay attention to these aspects can hamper feasibility, as well as further uptake and impact of diagnostics.

Proposed evidence for 'Evidence to Decision' tables in GRADE

Patient values and preferences:

Patients in high-burden settings value reaching diagnostic closure (finally knowing what is wrong with me) comprising of a diagnosis that is accurate, a diagnosis that is timely and is not delayed (as this exacerbates existing financial hardships and emotional and physical suffering and make patients feel guilty for infecting others (especially children)), that happens at facilities that are accessible and in the process of which patients can keep diagnosis-associated costs (travel, missing work) low. (Moderate confidence, applicable to index test 1,2, 3)

Equity:

Lengthy diagnostic delays, underutilization of diagnostics, lack of TB diagnostic facilities at lower levels and too many eligibility restrictions, hamper access to prompt and accurate testing and treatment particularly for vulnerable groups. (High confidence, applicable to index test 1,2, 3)
The identified challenges with low-complexity automated NAAT's utilization and accumulated delays risk undoing the added value as identified by the users, ultimately leading to underutilization and hamper access to prompt and accurate testing and treatment particularly for vulnerable groups**.** (High confidence, applicable to index test 1)
Staff and managers voiced concerns regarding sustainability of funding and maintenance, complex conflicts of interest between donors and implementers and concerns related to the strategic and equitable use of resources, which negatively affects creating equitable access to low-complexity automated NAATs. (High confidence, applicable to index test 1,2, 3)

Acceptability:

Patients can be reluctant to test for TB/MDR-TB because of stigma related to MDR-TB or related to having interrupted treatment in the past, because of fears of side effects, the failure to recognize symptoms, the inability to produce sputum and the cost, distance and travel concerns related to (repeat) clinic visits. (High confidence, applicable to index test 1,2, 3)
Health workers can be reluctant to test for TB or MDR-TB because of TB associated stigma and consequences for their patients, fears of acquiring TB, fear from supervisors when reclassifying patients already on TB treatment who turn out to be misclassified, fear of side effects of drugs in children, and community awareness of disease manifestations in children. (High confidence, applicable to index test 1,2, 3)
Compared to existing tests/sputum microscopy, **health workers** appreciate the rapidity and accuracy of low-complexity automated NAAT results, the confidence that this generates to start treating and motivate patients, the diversity of sample types, the ability to detect drug resistance (earlier or at all, for as many drugs as possible and altering clinician's risk perception of drug resistance in children), as well as the consequence of avoiding costlier investigations or hospital stays when using low-complexity automated NAAT. (High confidence, applicable to index test 1)
Laboratory technicians appreciate the improvement of overall laboratory work that low-complexity automated NAAT brings compared to sputum microscopy in terms of ease of use, ergonomics, and biosafety. (High confidence, applicable to index test 1)
Laboratory managers appreciate that monitoring of laboratory work and training is easier than with sputum microscopy and that low-complexity automated NAAT eases staff retention, as it increases staff

satisfaction and has a symbolic meaning of progress within the TB world. (Low confidence, applicable to index test 1)
The identified feasibility challenges with low-complexity automated NAAT utilization and accumulated delays at every step risk undoing the added value/benefits as identified by the **users** (avoiding delays, keeping cost lost, accurate results, drug resistant information, easing laboratory work), ultimately leading to underutilization (High confidence, applicable to index test 1). We can assume that if these values are not met users are less likely to find low-complexity automated NAATs acceptable.

Feasibility:

Feasibility is challenged by accumulation of diagnostic delays and/or underutilization at every step due to mainly health system factors: non-adherence to testing algorithms, testing for (MDR)-TB late in the process, empirical treatment, false negatives due to technology failure, large sample volumes and staff shortages, poor/delayed sample transport and sample quality, and result communication, delays in scheduling follow up visits and recalling patients, inconsistent result recording; lack of sufficient resources and maintenance (i.e. stock-outs; unreliable logistics; lack of funding, electricity, space, air conditioners, and sputum containers; dusty environment, and delayed or absent local repair option); inefficient/unclear work- and patient flows (for instance inefficient organizational processes, poor links between providers, unclear follow up mechanisms or where patients need to go); and lack of data-driven and inclusive national implementation processes. These challenges lead to delays and/or underutilization. (High confidence, applicable to index test 1,2,3)
Implementation of new diagnostics must be accompanied with training for clinicians, to help them interpret results from new molecular tests and understand how this relates to treatment of a patient (applicable to index test 1,2,3). In the past, with introduction of low-complexity automated NAAT this has been a challenge leading to underutilization (High confidence) or overreliance on low-complexity automated NAAT results at the expense of clinical acumen (Moderate confidence).
Poor sample quality, inconvenient sample collection facilities, non-functioning sample transport mechanisms, and difficulty of obtaining pediatric samples can cause error results and underutilization of low-complexity automated NAATs**.** (High confidence, applicable to index test 1)
Low-complexity automated NAAT seems to decrease workload in the laboratory in terms of freeing up time for laboratory staff, but in most settings the introduction of low-complexity automated NAAT increases workload of laboratory staff if added onto existing work without adjusting staffing arrangements, or if it does not replace existing diagnostic tests with the result that staff may be hesitant to accept testing with low-complexity automated NAAT**.** (Moderate confidence, applicable to index test 1)

Reflections during the GDG on qualitative evidence from the GDG members:
During the GDG an engaged discussion took place on the value of the qualitative evidence (qualitative evidence synthesis and interview study presented). They key points were:

- GDG panel members thought that evidence on the perspectives of communities, patients, different cadres of healthcare workers (including nurses, community health workers for contract tracing; laboratory staff and clinicians) is valuable and important input during a GDG.
- They were interested to hear different perspectives on uncertainty regarding diagnostic results which is often tricky to apply in clinical practice and on the work done by clinicians, laboratory staff and patients to make sense of diagnostic results. They realized that improved training for nurses and clinicians on diagnostics, including training visits in laboratories, will improve decisions to order tests and also communication with patients.

- They suggested adding qualitative methods routinely and from the beginning to diagnostic evaluation and clinical studies.
- One member suggested exploring collaborations with costing studies.

Acknowledgements

We are grateful to Vittoria Lutje, Cochrane Infectious Diseases' Information Specialist, for helping us with the search strategy. We are grateful to Darrel Ronald for helping us with the design of the figure.

References

Albert, H., Nathavitharana, R. R., Isaacs, C., Pai, M., Denkinger, C. M., & Boehme, C. C. (2016). Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? [10.1183/13993003.00543-2016]. *European Respiratory Journal, 48*, 516-525.

Bainomugisa, A., Gilpin, C., Coulter, C., & Marais, B. J. (2020). New Xpert MTB/XDR: added value and future in the field. *European Respiratory Journal, 56*(5), 2003616.

Beisel, U., Umlauf, R., Hutchinson, E., & Chandler, C. I. (2016). The complexities of simple technologies: re-imagining the role of rapid diagnostic tests in malaria control efforts. *Malaria Journal, 15*, 64.

Cattamanchi, A., Berger, C. A., Shete, P. B., Turyahabwe, S., Joloba, M., Moore, D. A. J., et al. (2020). Implementation science to improve the quality of tuberculosis diagnostic services in Uganda. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases, 18*, 100136.

Colvin, C. J., Leon, N., Wills, C., van Niekerk, M., Bissell, K., & Naidoo, P. (2015). Global-to-local policy transfer in the introduction of new molecular tuberculosis diagnostics in South Africa. *The International Journal of Tuberculosis and Lung Disease, 19*(11), 1326-1338.

Conradie, F., Diacon, A. H., Ngubane, N., Howell, P., Everitt, D., Crook, A. M., et al. (2020). Treatment of Highly Drug-Resistant Pulmonary Tuberculosis. *New England Journal of Medicine, 382*(10), 893-902.

Covidence. (2017). Covidence systematic review software. Melbourne: Veritas Health Innovation.

Creswell, J., Codlin, A. J., Andre, E., Micek, M. A., Bedru, A., Carter, E. J., et al. (2014). Results from early programmatic implementation of Xpert MTB/RIF testing in nine countries. *BMC Infectious Diseases, 14*(1), 2.

Di Tanna, G. L., Khaki, A. R., Theron, G., McCarthy, K., Cox, H., Mupfumi, L., et al. (2019). Effect of Xpert MTB/RIF on clinical outcomes in routine care settings: individual patient data meta-analysis. *Lancet Glob Health, 7*(2), e191-e199.

Engel, N., Davids, M., Blankvoort, N., Pai, N. P., Dheda, K., & Pai, M. (2015). Compounding diagnostic delays: a qualitative study of point-of-care testing in South Africa. *Tropical medicine & international health, 20*(4), 493-500.

Engel, N., Wachter, K., Pai, M., Gallarda, J., Boehme, C., Celentano, I., et al. (2016). Addressing the challenges of diagnostics demand and supply: insights from an online global health discussion platform. *BMJ Global Health, 1*(4).

England, K., Masini, T., & Fajardo, E. (2019). Detecting tuberculosis: rapid tools but slow progress. *Public Health Action, 9*(3), 80-83.

Eshun-Wilson, I., Rohwer, A., Hendricks, L., Oliver, S., & Garner, P. (2019). Being HIV positive and staying on antiretroviral therapy in Africa: A qualitative systematic review and theoretical model. *PLOS ONE, 14*(1), e0210408.

Haraka, F., & al., E. (2019). *Impact of diagnostic test Xpert MTB/RIF® on health outcomes for tuberculosis*. Paper presented at the World Health Organization-convened Guideline Development Group Meeting on "Molecular assays as initial tests for the diagnosis of tuberculosis and rifampicin resistance", 3-6 December 2019.

Hoang, T., Nguyen, N., Dinh, S., Nguyen, H., Cobelens, F., Thwaites, G., et al. (2015). Challenges in detection and treatment of multidrug resistant tuberculosis patients in Vietnam. *BMC Public Health, 15*(1), 980.

Jaroslawski, S., & Pai, M. (2012). Why are inaccurate tuberculosis serological tests widely used in the Indian private healthcare sector? A root-cause analysis. *Journal of Epidemiology and Global Health, 2*, 39-50.

Joshi, B., Lestari, T., Graham, S. M., Baral, S. C., Verma, S. C., Ghimire, G., et al. (2018). The implementation of Xpert MTB/RIF assay for diagnosis of tuberculosis in Nepal: A mixed-methods analysis. *PLOS ONE, 13*(8), e0201731.

Kelly-Cirino, C. D., Nkengasong, J., Kettler, H., Tongio, I., Gay-Andrieu, F., Escadafal, C., et al. (2019). Importance of diagnostics in epidemic and pandemic preparedness. *BMJ Global Health, 4*(Suppl 2), e001179.

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gülmezoglu, M., et al. (2015). Using Qualitative Evidence in Decision Making for Health and Social Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses (GRADE-CERQual). *PLOS Medicine, 12*(10), e1001895.

McDowell, A., & Pai, M. (2016). Treatment as diagnosis and diagnosis as treatment: empirical management of presumptive tuberculosis in India. *Int J Tuberc Lung Dis, 20*(4), 536-543.

McDowell, A., Raizada, N., Khaparde, S. D., Rao, R., Sarin, S., Kalra, A., et al. (2018). "Before Xpert I only had my expertise": A qualitative study on the utilization and effects of Xpert technology among pediatricians in 4 Indian cities. *PLOS ONE, 13*(3), e0193656.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med, 6*(7), e1000097.

Mwaura, M., Kao, K., Wambugu, J., Trollip, A., Sikhondze, W., Omesa, E., et al. (2020). Situating trade-offs: Stakeholder perspectives on overtreatment versus missed diagnosis in transition to Xpert MTB/RIF Ultra in Kenya and Swaziland. *PLOS ONE, 15*(2), e0228669.

Naidoo, P., van Niekerk, M., du Toit, E., Beyers, N., & Leon, N. (2015). Pathways to multidrug-resistant tuberculosis diagnosis and treatment initiation: a qualitative comparison of patients' experiences in the era of rapid molecular diagnostic tests. *BMC Health Services Research, 15*(1), 488.

Nalugwa, T., Shete, P. B., Nantale, M., Farr, K., Ojok, C., Ochom, E., et al. (2020). Challenges with scale-up of GeneXpert MTB/RIF® in Uganda: a health systems perspective. *BMC Health Services Research, 20*(1), 162.

Newtonraj, A., Venables, E., Selvaraj, K., Kundu, D., Purty, A. J., Manikandan, M., et al. (2019). Xpert negative means no TB: A mixed-methods study into early implementation of Xpert in Puducherry, India. *J Family Med Prim Care, 8*(4), 1379-1385.

Noyes, J., Booth, A., Cargo, M., Flemming, K., Harden, A., Harris, J., et al. (2020). Chapter 21: Qualitative evidence. In J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020)*: Cochrane.

Oliwa, J. N., Odero, S. A., Nzinga, J., van Hensbroek, M. B., Jones, C., English, M., et al. (2020). Perspectives and practices of health workers around diagnosis of paediatric tuberculosis in hospitals in a resource-poor setting - modern diagnostics meet age-old challenges. *BMC health services research, 20*(1), 708-708.

Pai, M., Schumacher, S. G., & Abimbola, S. (2018). Surrogate endpoints in global health research: still searching for killer apps and silver bullets? *BMJ Global Health, 3*(2).

Phyo, A. M., Kumar, A. M. V., Soe, K. T., Kyaw, K. W. Y., Thu, A. S., Wai, P. P., et al. (2019). Contact Investigation of Multidrug-Resistant Tuberculosis Patients: A Mixed-Methods Study from Myanmar. *Tropical medicine and infectious disease, 5*(1), 3.

Pooran, A., Theron, G., Zijenah, L., Chanda, D., Clowes, P., Mwenge, L., et al. (2019). Point of care Xpert MTB/RIF versus smear microscopy for tuberculosis diagnosis in southern African primary care clinics: a multicentre economic evaluation. *The Lancet Global Health, 7*(6), e798-e807.

R. de Camargo, K., R. Guedes, C., Caetano, R., Menezes, A., & Trajman, A. (2015). The adoption of a new diagnostic technology for tuberculosis in two Brazilian cities from the perspective of patients and healthcare workers: a qualitative study. *BMC Health Services Research, 15*(1), 275.

Rees, R. W., Caird, J., Dickson, K., Vigurs, C., & Thomas, J. (2014). 'It's on your conscience all the time': a systematic review of qualitative studies examining views on obesity among young people aged 12–18 years in the UK. *BMJ Open, 4*(4), e004404.

Rendell, N. L., Bekhbat, S., Ganbaatar, G., Dorjravdan, M., Pai, M., & Dobler, C. C. (2017). Implementation of the Xpert MTB/RIF assay for tuberculosis in Mongolia: a qualitative exploration of barriers and enablers. *PeerJ, 5*, e3567.

Royce, S., Khann, S., Yadav, R. P., Mao, E. T., Cattamanchi, A., Sam, S., et al. (2014). Identifying multidrug resistance in previously treated tuberculosis patients: a mixed-methods study in Cambodia. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 18*(11), 1299-1306.

Shewade, H. D., Kokane, A. M., Singh, A. R., Parmar, M., Verma, M., Desikan, P., et al. (2018). Provider reported barriers and solutions to improve testing among tuberculosis patients 'eligible for drug susceptibility test': A qualitative study from programmatic setting in India. *PLOS ONE, 13*(4), e0196162.

Stime, K. J., Garrett, N., Sookrajh, Y., Dorward, J., Dlamini, N., Olowolagba, A., et al. (2018). Clinic flow for STI, HIV, and TB patients in an urban infectious disease clinic offering point-of-care testing services in Durban, South Africa. *BMC Health Serv Res, 18*(1), 363.

Vijayageetha, M., Kumar, A. M., Ramakrishnan, J., Sarkar, S., Papa, D., Mehta, K., et al. (2019). Tuberculosis screening among pregnant women attending a tertiary care hospital in Puducherry, South India: is it worth the effort? *Global health action, 12*(1), 1564488-1564488.

World Health, O. (2015). *Implementing the end TB strategy: the essentials*. Geneva: World Health Organization.

World Health Organization. (2020a). *Global Tuberculosis Report 2020* Geneva: World Health Organization.

World Health Organization. (2020b). *WHO consolidated guidelines on tuberculosis. Module 3: Diagnosis - Rapid diagnostics for tuberculosis detection*. Geneva: World Health Organization.

World Health Organization. (2020c). *WHO consolidated guidelines on tuberculosis. Module 4: treatment* Geneva: World Health Organization.

Appendices

Appendix 1: Search strategy

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R) <1946 to September 05, 2020>
Search Strategy:
---

1 Extensively Drug-Resistant Tuberculosis/ or Tuberculosis/ or tuberculosis.mp. or Tuberculosis, Multidrug-Resistant/ or Tuberculosis, Pulmonary/ or Mycobacterium tuberculosis/
2 (Tuberculosis or MDR-TB or XDR-TB or tuberculous).ti. or (Tuberculosis or MDR-TB or XDR-TB or tuberculous).ab.
3 1 or 2
4 (Truenat or Cepheid or Xpert*).mp.
5 Genexpert*.mp.
6 drug susceptibility test*.mp.
7 (cartridge adj3 test*).mp.
8 cartridge*.ab. or cartridge*.ti.
9 exp Point-of-Care Systems/
10 Reagent Kits, Diagnostic/
11 Max MDR-TB assay.mp.
12 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11
13 3 and 12
14 "Patient Acceptance of Health Care"/ or acceptability.mp.
15 Health Equity/ or equity.mp. or Health Services Accessibility/
16 Patient Preference/ or preferences.mp.
17 Patient Satisfaction/ or Attitude to Health/
18 barriers.mp.
19 challenges.mp.
20 patient experience*.mp.
21 "Attitude of Health Personnel"/ or providers experience*.mp.
22 Critical Pathways/
23 facilitator*.ab. or facilitator*.ti.
24 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23
25 13 and 24
26 Interviews as Topic/ or interview*.mp. or Interview/
27 survey*.mp. or Health Surveys/ or Health Care Surveys/ or "Surveys and Questionnaires"/
28 Qualitative Research/
29 Focus group discussion*.mp. or Focus Groups/
30 "mixed methods".ti. or "mixed methods".ab. or "mixed-methods".ti. or "mixed-methods".ab.
31 26 or 27 or 28 or 29 or 30
32 13 and 31
33 25 or 32
34 limit 33 to yr="2007 -Current"
35 "systematic review"/
36 (metaanalysis or meta-analysis).mp.
37 35 or 36
38 13 and 37
39 limit 38 to yr="2007 -Current"

Database: Embase <1996 to 2020 Week 36>
Search Strategy:
---

1 tuberculosis/ or tuberculosis.mp.
2 drug resistant tuberculosis.mp. or drug resistant tuberculosis/
3 multidrug resistant tuberculosis.mp. or multidrug resistant tuberculosis/
4 MDR-TB.mp.
5 XDR-TB.mp.
6 extensively drug resistant tuberculosis/
7 mycobacterium tuberculosis.mp. or Mycobacterium tuberculosis/
8 1 or 2 or 3 or 4 or 5 or 6 or 7
9 (Truenat or Cepheid or Xpert*).mp.
10 Genexpert*.mp.
11 drug susceptibility test*.mp.
12 (cartridge adj3 test*).mp.
13 cartridge*.ab. or cartridge*.ti.
14 "point of care testing"/
15 *diagnostic test/
16 diagnostic test accuracy study/
17 Max MDR-TB assay.mp.
18 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17
19 8 and 18
20 patient acceptance of care.mp. or patient attitude/
21 acceptability.mp.
22 patient preference/ or patient preference*.mp.
23 health equity.mp. or health equity/
24 Health Services Accessibility.mp. or health care access/
25 patient satisfaction.mp. or patient satisfaction/
26 barriers.mp.
27 challenges.mp.
28 patient experience*.mp.
29 Attitude of Health Personnel.mp. or health personnel attitude/
30 Critical Pathways.mp. or clinical pathway/
31 facilitator*.ab. or facilitator*.ti.
32 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31
33 19 and 32
34 Diagnostic Interview Schedule/ or exp interview/ or interview*.mp.
35 health care survey/ or survey*.mp. or health survey/
36 (Surveys and Questionnaires).mp.
37 qualitative research.mp. or qualitative research/
38 focus group.mp.
39 (mixed adj2 method*).mp.
40 34 or 35 or 36 or 37 or 38 or 39
41 19 and 40
42 33 or 41
43 limit 42 to yr="2007 -Current"
44 systematic review.mp. or "systematic review"/
45 meta analysis/
46 metaanalysis.mp.
47 44 or 45 or 46
48 19 and 47

49 limit 48 to yr="2007 -Current"

Cinahl (EBSCOHost) Limiters - Published Date: 20070101-20201231

#	Query
S22	S8 AND S20
S21	S8 AND S20
S20	TX systematic review or meta-analysis
S19	S14 OR S18
S18	S8 AND S17
S17	S15 OR S16
S16	TX focus group*
S15	TX interview* OR TX (survey* or questionnaire*) OR TX (qualitative research or qualitative study or qualitative methods or mixed methods)
S14	S8 AND S12
S13	S8 AND S12
S12	S9 OR S10 OR S11
S11	TX (barriers or challenges) OR TX critical pathway OR TX facilitator*
S10	TX patient preference* OR TX (patient satisfaction or patients experiences or patients perceptions or patients attitudes)
S9	TX acceptance of care OR TX health equity OR MW Health Services Accessibility
S8	S3 AND S7
S7	S4 OR S5 OR S6
S6	TX Max MDR-TB OR TI cartridge OR AB cartridge
S5	TX drug susceptibility test* OR TX cartridge N2 test* OR TX point of care testing
S4	TX Truenat or Cepheid or Xpert* or Genexpert*
S3	S1 OR S2
S2	TX extensively drug resistant tuberculosis OR MH tuberculosis, multidrug-resistant
S1	TX (tuberculosis or TB or MDR-TB or XDR-TB) OR MW mycobacterium tuberculosis OR MW multidrug resistant tuberculosis

PsycInfo (EBSCOHost) Limiters - Published Date: 20070101-20201231

#	Query
S22	S8 AND S20
S21	S8 AND S20
S20	TX systematic review or meta-analysis
S19	S14 OR S18
S18	S8 AND S17
S17	S15 OR S16
S16	TX focus group*

S15	TX interview* OR TX (survey* or questionnaire*) OR TX (qualitative research or qualitative study or qualitative methods or mixed methods)
S14	S8 AND S12
S13	S8 AND S12
S12	S9 OR S10 OR S11
S11	TX (barriers or challenges) OR TX critical pathway OR TX facilitator*
S10	TX patient preference* OR TX (patient satisfaction or patients experiences or patients perceptions or patients attitudes)
S9	TX acceptance of care OR TX health equity OR Health Services Accessibility
S8	S3 AND S7
S7	S4 OR S5 OR S6
S6	TX Max MDR-TB OR TI cartridge OR AB cartridge
S5	TX drug susceptibility test* OR TX cartridge N2 test* OR TX point of care testing
S4	TX Truenat or Cepheid or Xpert* or Genexpert*
S3	S1 OR S2
S2	TX extensively drug resistant tuberculosis OR tuberculosis, multidrug-resistant
S1	TX (tuberculosis or TB or MDR-TB or XDR-TB) OR mycobacterium tuberculosis OR multidrug resistant tuberculosis

Appendix 2: Assessment of Methodological Limitations

Study ID	Were steps taken to increase rigour in the sampling?	Were steps taken to increase rigour in the data collected?	Were steps taken to increase rigour in the analysis of the data?	Were the findings of the study grounded in/ supported by the data?	Please rate the findings of the study in terms of their breadth and depth.
McDowell 2016	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, several steps were taken	Yes, a fairly thorough attempt was made	Yes, several steps were taken
McDowell 2018	Yes, a fairly thorough attempt was made	Yes, a few steps were taken	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, several steps were taken
Naidoo 2015	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, several steps were taken	Yes, a fairly thorough attempt was made	Yes, several steps were taken
Engel 2015	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made
Oliwa 2020	Yes, several steps were taken	Yes, several steps were taken	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made	Yes, a fairly thorough attempt was made
deCamargo 2015	Yes, several steps were taken	Yes, several steps were taken	Yes, several steps were taken	Yes, a fairly thorough attempt was made	Yes, several steps were taken
Colvin 2015	Yes, several steps were taken	Yes, several steps were taken	Yes, several steps were taken	Yes, a few steps were taken	Yes, several steps were taken
Shewade 2018	Yes, several steps were taken	Yes, several steps were taken	Yes, a fairly thorough attempt was made	Yes, several steps were taken	Yes, a few steps were taken
Mwaura 2020	Yes, several steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	Yes, several steps were taken	Yes, several steps were taken
Rendell 2017	Yes, a few steps were taken	Yes, several steps were taken	Yes, several steps were taken	Yes, a few steps were taken	Yes, a few steps were taken
Royce 2014	Yes, several steps were taken	Yes, several steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken
Phyo 2019	Yes, a few steps were taken	Yes, a few steps were taken	Yes, several steps were taken	Yes, a few steps were taken	Yes, a few steps were taken
Joshi 2018	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken
Hoang 2015	Yes, a few steps were taken	Yes, several steps were taken	No, not at all/Not stated/ Can't tell	Yes, a few steps were taken	Yes, a few steps were taken
Newtonraj 2019	Yes, a few steps were taken	Yes, a few steps were taken	Yes, several steps were taken	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell
Vijayageetha 2019	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken
Creswell 2014	Yes, a few steps were taken	Yes, a few steps were taken	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell	Yes, several steps were taken
Stime 2018	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell	Yes, a few steps were taken	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell
Jaroslawski 2012	Yes, several steps were taken	No, not at all/Not stated/ Can't tell	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell	No, not at all/Not stated/ Can't tell
Cattamanchi 2020	No, not at all/Not stated/ Can't tell	No, not at all/Not stated/ Can't tell	No, not at all/Not stated/ Can't tell	No, not at all/Not stated/ Can't tell	No, not at all/Not stated/ Can't tell
Nalugwa 2020	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell	Yes, a few steps were taken	No, not at all/Not stated/ Can't tell	No, not at all/Not stated/ Can't tell

Appendix 3: CERQual summary of findings

Finding #	Review finding	Methodological limitations	Coherence	Relevance	Adequacy	CERQual assessment of confidence in the evidence	Explanation of CERQual assessment	Studies contributing to review finding
Critical aspects users value								
1	**Patients** in high-TB burden countries value 1) getting an accurate diagnosis and reaching diagnostic closure (finally knowing what is wrong with me), 2) avoiding diagnostic delays as they exacerbate existing financial hardships and emotional and physical suffering and make patients feel guilty for infecting others (especially children), 3) having accessible facilities and 4) reducing diagnosis-associated costs (travel, missing work) as important outcomes of the diagnostic.	minor concerns - across the four components of the methodological limitations tool, three of the studies contributing took a few steps to ensure methodological quality and the remaining three took several steps to ensure methodological quality	no concerns - synthesis was directly related to primary studies, missing explanations were explored and added to the finding	minor concerns about study locations: studies mostly located in urban areas in high burden settings, good variety of facility types	minor concerns- two rich studies included and the additional four all have undertaken few steps towards richness; the number of participants included is adequate for qualitative designs	moderate confidence	we have minor concerns about methodological quality and adequacy and we have minor concerns about relevance (because of its importance to the finding)	Naidoo 2015, deCamargo 2015, Joshi 2018, Phyo 2020, Royce 2014, Vijayageetha 2019
2	Compared to existing tests/sputum microscopy, **healthcare professionals** appreciate the rapidity and accuracy of low-complexity automated NAAT results, the diversity of sample types, ability to detect drug resistance, as well as the consequence of avoiding costlier investigations or hospital stays when using low-complexity automated NAATs..	minor concerns - across the four components the methodological quality was fairly high for three studies and the remaining studies took mostly a few steps ti increase quality	no concerns - synthesis was directly related to primary studies	no concerns -good variety of facilities, public/private and type of healthcare workers, and fairly diverse set of countries, studies mostly located in urban areas in high burden settings but we do not think this would affect relevance for this finding	minor - three rich studies included, only one thin study and the rest have undertaken several steps towards richness, number of participants included is adequate	high confidence	mainly because we have no concerns about coherence and relevance and only minor concerns about methodological quality and richness of a few studies	Rendell 2017, Newtonraj 2020, Joshi 2018, Mwaura 2020, McDowell 2018, deCamargo 2015,Vijayageetha 2019, Naidoo 2015
3	Low-complexity automated NAAT allows **healthcare workers** to detect drug resistance earlier and pediatricians in particular mentioned how it heightened their risk perception of drug resistance in children; yet in a context with widespread severe forms of drug resistance and a habit of treating empirically first, **clinicians** see the inability to detect resistance of some NAATs beyond rifampicin as a hindrance..	no concerns - the majority of the studies were of good quality	minor concerns - good fit of finding with primary studies, but change in risk perception in need for entire resistance profile mentioned in only one study each, but these were studies well grounded in the data	no/very minor concerns, countries with large DR burden included, except examples from Eastern Europe missing, public/private, urban/rural, good variety of primary care and CB-NAAT testing centre facilities	no concerns - three rich studies of four, adequate numbers of participants	high confidence	mainly because quality of studies is high and we only have a minor concern about coherence due to number of studies contributing to each part of the finding	McDowell 2018, McDowell 2016, Naidoo 2015, Joshi 2018, deCamargo 2015
4	**Clinicians** value the confidence that low-complexity automated NAAT results generate, to start treatment, to reassure and motivate patients and their caretakers, to justify actions towards other doctors and to increase collaboration between private/public providers.	no concerns - the studies were of good quality	no concerns - good fit of finding with primary studies, other explanations of how confidence matters are captured in finding #13	minor concerns because it is only two countries, but good variety of participants, facilities and public/private providers	no/very minor concerns - just two but very rich studies with adequate numbers of participants	confidence	we have no concerns or very minor concerns across all components	McDowell 2018, Oliwa 2020
5	**Laboratory technicians** appreciate the improvement of overall laboratory work that low-complexity automated NAAT brings compared to sputum microscopy in terms of ease of use, ergonomics, and biosafety	minor concerns - one study of good quality, the remaining two took a few steps to ensure methodological quality	no concerns - good fit of finding with primary study	no concerns - variety of locations, countries, facilities and users included	minor concerns- two relatively rich studies, the third one is not so rich, the number and type of participants included is adequate	high confidence	we have no concerns or very minor concerns across all components	deCamargo 2015, Creswell, 2014, Newtonraj 2020

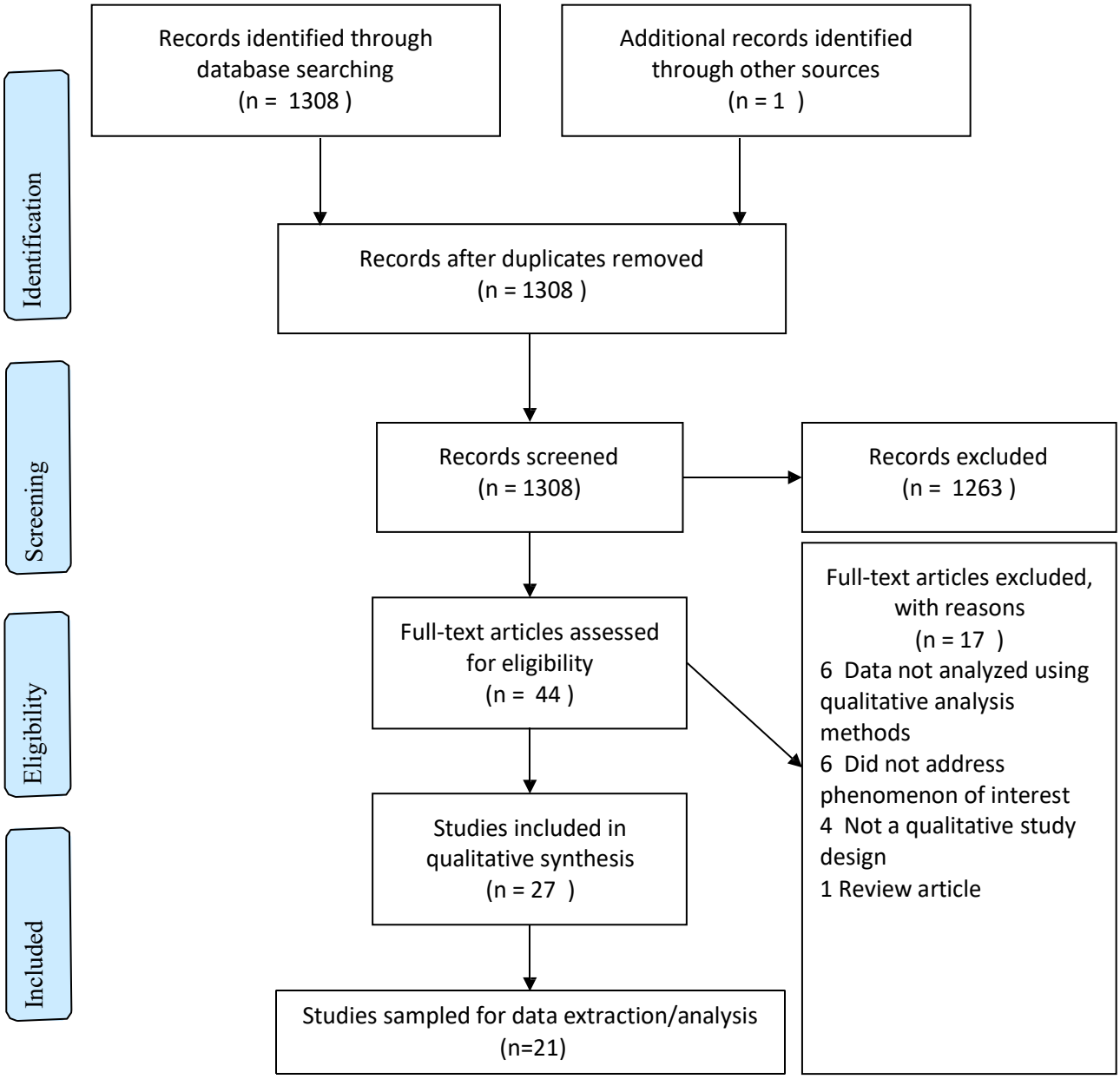
6	**Laboratory managers** appreciate that monitoring of laboratory work and training is easier than with sputum microscopy and that low-complexity automated NAAT eases staff retention, as it increases staff satisfaction and has a symbolic meaning of progress within the TB world..	no/very minor concerns	no concerns - good fit of finding with primary study	serious concerns - just one setting (urban, public clinic), study early in implementation of CB-NAAT	moderate concerns - because it is only one study, but it is rich, with an adequate number of participants but unclear how many of these were managers	low confidence	we have serious or moderate concerns about adequacy and relevance and no concerns about methodological quality and coherence	deCamargo 2015

Finding #	Review finding	Methodological limitations	Coherence	Relevance	Adequacy	CERQual assessment of confidence in the evidence	Explanation of CERQual assessment	Studies contributing to review finding
Challenges to realizing these values								
7	**Patients can be reluctant to test** for TB/MDR-TB because of stigma related to MDR-TB or related to having interrupted treatment in the past, because of fears of side effects, the failure to recognize symptoms, the inability to produce sputum and the cost, distance and travel concerns related to (repeat) clinic visits.	no/very minor concerns- three out of four included studies have fairly good methodological quality across all four components	no concerns - good fit of finding with primary study	no/very minor concerns, varied participants, facilities, urban/rural, even though just four countries but we do not expect adding more countries would have altered finding substantially	minor concerns - one rich study and the remaining took a few steps towards richness, the number of participants is adequate	high confidence	we have no concerns or very minor concerns across all components	Shewade 2018, Phyo 2019, Royce 2014, Naidoo 2015
8	**Health workers can be reluctant to test** for TB or MDR-TB because of TB associated stigma and consequences for their patients, fears of acquiring TB, fear from supervisors when reclassifying patients already on TB treatment who turn out to be misclassified, fear of side effects of drugs in children, and community awareness of disease manifestations in children.	no/very minor concerns - one study of very high quality, the other took several steps towards high quality	no concerns - good fit of finding with primary study	no concerns - variety of facilities, participants, not the usual dominantly represented countries and at two different time points	no/very minor concerns, one rich study and one study which seemed rich but quotes were not attributable, might have been just a reporting issue; very adequate numbers of participants for both studies	high confidence	we have no concerns across all components	Oliwa 2020, Royce 2014
9	**Diagnostic delays are accumulated** because of various health system factors (i.e. non-adherence to testing algorithms, testing for (MDR-)TB late in the process, empirical treatment, false negatives due to technology failure, large sample volumes and staff shortages, poor/delayed sample transport and result communication, delays in scheduling follow up visits and recalling patients, inconsistent result recording) and to a lesser extent patient-related delays (i.e. missed follow-up appointments, competing family demands and seeking traditional health-care).	minor concerns - varied methodological quality of included studies, three of high quality, studies of lower quality do not contribute new or additional insights, rather confirm other studies	no concerns - descriptive and specific statement based on the data from primary study	no concerns - good variety of users, facilities, public/private, urban/rural, time points and countries	minor concerns, four relatively rich studies, adequate numbers of participants, well known descriptive finding	high confidence	we have no or very minor concerns across the components also because diagnostic delay is well established and half the studies are very rich studies with a great relevance and then weaker studies findings point into the same direction	McDowell 2016,Naidoo 2015, Nalugwa 2020, Cattamanchi 2020, Engel, 2015, Stime 2018, , Rendell 2017, Creswell 2014, Royce 2014
10	Poor **sample** quality, inconvenient sample collection facilities, non-functioning sample transport mechanisms, and difficulty of obtaining pediatric samples can cause error results and underutilization of low-complexity automated NAAT.	minor concerns - of 12 studies contributing to the finding, about half were of good quality while the other half took a few steps towards methodological quality	no concerns - descriptive and specific statement based on the data from primary study	no concerns - good variety of users, facilities, public/private, urban/rural settings, time points and countries	minor concerns, three rich studies and the others took a few steps towards richness, adequate number of participants	high confidence	mainly because we have no concerns about coherence and relevance and only minor concerns about the methodological quality of half the studies contributing and no concerns about the quality and richness of the remaining ones	Rendell 2017, Royce 2014, Newtonraj 2019, Hoang, 2018, Vijayageetha 2019, Shewade 2018, Phyo 2019, Creswell 2014, Cattamanchi 2020, McDowell 2016, McDowell 2018, Oliwa 2020
11	The **lack of sufficient resources and of ensuring maintenance** (i.e. stock-outs; unreliable logistics; lack of funding, electricity, space, air conditioners, and sputum containers; dusty environment, and delayed or absent local repair option) leads to higher test failure rates and underutilization and negatively influences uptake and impact of low-complexity automated NAAT.	minor concerns - of 10 studies contributing to the finding, about 5 were of fairly good or very quality while the other half took a few steps towards methodological quality	no concerns - captures the data from primary studies	no/very minor concerns, good variety of users, facilities, countries, urban/rural, time points, the majority of studies in public sector settings	minor concerns - four rich studies and the others took a few steps towards richness while for two studies it was not clearly reported, adequate number of participants	high confidence	mainly because we have no concerns about coherence and relevance and only minor concerns about methodological quality and richness of about half the studies	Rendell 2017, Oliwa 2020, Hoang 2015, Nalugwa 2020, Joshi 2018, Mwaura 2020, deCamargo 2015, Stime 2018, Creswell 2014, Shewade 2018, Hoang 2015,

Finding #	Review finding	Methodological limitations	Coherence	Relevance	Adequacy	CERQual assessment of confidence in the evidence	Eplanation of CERQual assessment	Studies contributing to review finding
12	Low-complexity automated NAAT seems to decrease workload by freeing up time for laboratory staff, but in most settings staff may be hesitant to accept testing with low-complexity automated NAAT because it increases workload if added onto existing laboratory work without adjusting staffing arrangements, or if it does not replace existing diagnostic tests.	minor concerns - of the 7 studies contributing, the majority of studies took a few steps towards methodological quality with 3 taking several steps	minor concerns - finding captures the primary studies well, the only minor concern is that the explicit mentioning of not accepting CB-NAAT because of workload concerns was only mentioned in one study	no/very minor concerns, good variety of users, facilities, and countries, though predominantly urban and public facilities (but that is expected for this finding because of where CB-NAAT is mainly located)	minor concerns - the studies took a few steps towards richness and had an adequate number of participants	moderate confidence	mainly because of the minor concern with coherence where only one study contributed to the point on acceptance	deCamargo 2015, Joshi 2018, Rendell 2017, Phyo, 2019, Shewade 2018, Stime 2018, Vijayageetha 2019
13	Workflows, professional roles and patient flows matter for utilizing low-complexity automated NAAT, for instance inefficient organizational processes, poor links between providers, unclear follow up mechanisms or where patients need to go can delay diagnostic processes..	no/very minor concerns - mainly because methodological limitations were minor and related to not being reported and two studies were well done, the other three took a few steps	no concerns - captures the data from primary studies	no/very concerns - good variety of users, facilities, countries, and time points, but all studies located in public settings however the coordination between public/private sector is covered in one study	minor concerns - the studies took a few steps, one did not report on richness, two were rich, adequate number of participants	high confidence	no concerns about methodological quality, coherence and relevance, we only have minor concerns about the degree of richness	Royce 2014, Oliwa 2020, Stime 2018, Hoang 2015, deCamargo 2015
14	Too much confidence in low-complexity automated NAAT's accuracy can mean blindly accepting results without using clinical impressions or for patients to trust low-complexity automated NAAT because it is a computer-based result.	moderate concerns - no study of very high quality, three studies took a few steps towards methodological quality	no concerns - captures the data from primary studies	no concerns- good variety of users, facilities, countries and time points, public/private, rural/urban	moderate concerns - because there is only one study that took several steps towards richness, the two remaining are thin or not reported, the number of participants is adequate	moderate confidence	mainly because of the moderate concerns with methodological quality and richness of data	Newtonraj 2019, Mwaura 2020, Joshi 2018
15	**Implementation processes** have been challenged by lack of data on pragmatic effectiveness in operational conditions, lack of knowledge and awareness among providers beyond lab personnel, lack of guidelines and standardized training modules and instructions and a lack of national policy consensus and inclusive decision-making prior to roll out.	minor concerns - 4 of 9 studies were of very good quality, the remaining ones took few or several steps towards increasing quality	no/very minor concerns - captures data from primary studies, the point on data and inclusive decision- making only made by Colvin but this is just a minor concern as the study is well grounded in data	no concerns- good variety of users, facilities, countries and time points, public/private, rural/urban	minor concerns - adequate number of participants, 4 of 9 studies took several steps towards richness, the others a few	high confidence	mainly because we have no concerns about coherence and relevance and only minor concerns about methodological quality and richness of about half the studies	Colvin 2015, Newtonraj 2019, Joshi, 2018, Rendell 2017, Shewade 2018, deCamargo 2015,Creswell 2014, Hoang 2015, Naidoo 2015

Finding #	Review finding	Methodological limitations	Coherence	Relevance	Adequacy	CERQual assessment of confidence in the evidence	Explanation of CERQual assessment	Studies contributing to review finding
Concerns for access/equity								
16	Staff and managers voiced concerns regarding sustainability of funding and maintenance, complex conflicts of interest between donors and implementers and concerns related to the strategic and equitable use of resources, which negatively affects creating equitable access to automated NAATs of low complexity.	no concerns - 2 out of 3 studies of good quality	no/very minor concerns - captures data from primary studies, the point on conflict of interest and strategic use of resources only made by Colvin but this is just a minor concern as the study is well grounded in data	no/very minor concerns- good variety of users, facilities, countries, public/private, rural/urban, only minor concern is that the studies are all from an early time point of Xpert implementation	minor concerns - they all are fairly rich, the number of implementer/manager participants is for two studies not clear, just three studies overall	high confidence	we have no concerns except minor concerns because part of the finding relies on only one study	Colvin 2015, deCamargo 2015, Cresswell 2014, Jaroslawski 2015
17	Lengthy diagnostic delays, underutilization of automated NAATs of low complexity, lack of TB diagnostic facilities at lower levels and too many eligibility restrictions, hamper access to prompt and accurate testing and treatment particularly for vulnerable groups.	minor concerns - 5 out of 11 studies are of very good methodological quality, the remaining ones took a few steps to increase quality across the assessed domains	no concerns - captures the data from primary studies and refers to summary findings #7 and 9	no concerns- good variety of users, facilities, countries and time points, public/private, rural/urban	minor concerns - 5 rich studies 4 took a few steps towards richness, , adequate numbers of participants	high confidence	we have only very minor concerns about methodological quality and richness of half the studies	Nalugwa 2020, McDowell 2018, Oliwa 2020, Joshi 2018, Naidoo 2015, Hoang 2015, McDowell 2016, Newtonraj 2019, Phyo 2019, Royce 2014, Engel 2015
18	The identified challenges and accumulated delays risk undoing the added value as identified by the users, ultimately leading to underutilization and important implications for access and equity	no concerns - high quality of included studies	no concerns - captures the data from the four primary studies and from the summary findings #1-15 which all were judged to be coherent except with two where we had minor concerns,	no concerns because it relates to summary finding #7-15 which have no/very minor concerns about relevance, the four directly contributing studies have good variety of users, facilities, urban/rural, public/private even if only focused on two countries	no concerns - rich studies and the large number of statements #7-15 contributing to this finding	high confidence	no concerns	McDowell 2018, Shewade 2018, Engel 2015. Naidoo 2015 (see memo dealys) (and this analysis)

Appendix 4: Figure 1: PRISMA flow chart



Appendix 5: Study characteristics table

First author	Year of publication	Country (income classification)	Geographical setting	Type of health facility	public/private	Background prevalence	Diagnostic technology	Programmatic features of the intervention (Where and how)	Target population	Total number of participants and types of settings	Research questions/ objectives	Data collection methods
Mwaura	2020	Kenya (lower middle income) and Eswatini (lower middle income)	Unclear/not reported	unclear/ unreported	unclear/not reported	High burden	Xpert MTB/RIF Ultra	Unclear	presumptive TB patients	47	to examine the views and norms of multiple TB stakeholders on the trade- off between overtreatment versus under diagnosis of TB, and to understand the role qualitative research can play in engaging in- country stakeholders during the launch and roll-out of new TB diagnostics	Focus group discussions (FGDs)
Cattamanchi	2020	Uganda (low income)	Both rural and urban	Health centers	Public facility	High burden	Xpert MTB/RIF	Testing sites (i.e., hubs), present in most districts of the country, are linked with 3–5 peripheral microscopy units (i.e., spokes) where sputum samples are collected and transported to the testing hubs. The results are returned to the microscopy centers. Intervention: Daily sputum transport to Xpert testing hubs was selected to facilitate same-day (or next-day) Xpert testing for all smear-negative patients.	presumptive TB patients	not clear	To identify key reasons at multiple levels for attrition along the TB diagnostic evaluation cascade of care. (within a larger mixed-method implementation research)	consultation with stakeholders and literature review
Nalugwa	2020	Uganda (low income)	Both rural and urban	Community health centres (clinics) and Xpert testing sites	Public facility	High burden	Xpert MTB/RIF	Uganda adopted policy recommendations in line with WHO guidelines; use of smear microscopy and Xpert MTB/RIF at participating health centers	presumptive TB patients presenting to community health facilities linked with TBdiagnostic units have access to rapid, referral-based Xpert testing. At the time of this study, Uganda national guidelines called for Xpert testing in persons living with HIV, health care workers, contacts of drug-resistant (DR-TB) patients, pregnant women or breast-feeding mothers, prisoners, patients from refugee camps, and diabetics.	N=23 participating community health centres (clinic staff)	mixed method, qualitative part: to assess the process of specimen collection, specimen transport, specimen testing, result reporting and patient linkage to treatment initiation if diagnosed with TB	Qualitative data was collected from field notes taken by study staff during site visits. Staff recorded observations about the TB diagnostic evaluation process in participating community health centers during site visits for trainings, surveys, and data abstraction.
Oliwa	2020	Kenya (lower middle income)	Unclear/not reported	County hospitals (Level 3)	Public facility	High burden	Xpert MTB/RIF assay	upfront Xpert for the diagnosis of pediatric TB in Kenyan county referral hospitals	Children	N=40 (29 interviews with front line health workers and mid-level managers. 3 small group discussions (N=6) and 5 key informant interviews with policy makers and senior health service administrative staff (medical officers; clinical officers; nursing officers; medical officer interns; clinical officer interns; nursing officer interns and laboratory technologists.)	to understand how context influences/shapes TB case detection and use of TB diagnostic tests including Xpert in children within hospitals	face-to-face semi-structured and key informant interviews, small group discussions, and observations of child TB trainings, sensitisation meetings and policy meetings, and hospital practices as well as desk review of relevant guidelines, job aides and policy documents.
Vijayageetha	2019	India (lower middle income)	Urban	Tertiary care hospital	Public facility	High burden	Xpert MTB/RIF	symptom screening and if positive then sputum and culture and per discretion of chest physician in case of high index of suspicion also low-complexity automated NAAT	pregnant women with TB symptoms, Xpert is mainly used for diagnosis of paediatric TB, HIV- associated TB, extrapulmonary TB and MDR-TB	N=7 (administrator n=1, obstetricians n=2 chest physicians n=1, physician n=1, nursing officers n=2)	mixed method study, to examine implementation challenges of TB screening among pregnant women from the healthcare providers perspective	interviews and observations

First author	Year of publication	Country (income classification)	Geographical setting	Type of health facility	public/private	Background prevalence	Diagnostic technology	Programmatic features of the intervention (Where and how)	Target population	Total number of participants and tpes of	Research questions/ objectives	Data collection methods
Phyo	2019	Myanmar (lower middle income)	Urban	Community outreach in townships	Other: NGO led	High burden	Xpert MTB/RIF	Household contacts of MDR-TB patients with TB symptoms are investigated using Xpert MTB/RIF; but policy is followed poorly in NTP. Irrespective of symptoms or chest radiography findings, people who are able to produce a sputum specimen are investigated further using sputum microscopy and Xpert MTB/RIF. Contacts who are unable to produce sputum or those with negative sputum results, but shadows suggestive of TB on chest radiography, are referred for further clinical management. In some township TB centres, Xpert MTB/RIF and chest radiography are not available.	household contacts of MDR-TB patients	N=21 household contacts of MDR-TB patients n=8, healthcare providers n=13 (community volunteers, project nurses), project supervisor	mixed methods: To explore the barriers in implementing contact investigation from the perspective of household contacts and health care providers.	interviews
Newtonraj	2019	India (lower middle icome)	Unclear/not reported	Clinics, designated microscopy centres and hospitals	Both public and private	High burden	Xpert MTB/RIF	Xpert MTB/RIF was located at the IRL (intermediate reference laboratory), along with culture and LPA; district microscopy centres (mostly within district hospitals and medical colleges) would send samples of eligible patients	Xpert as the initial diagnostic test for HIV-associated TB, EPTB, and pediatric TB and as an add-on test for sputum microscopy–negative patients if chest radiography was suggestive of TB.	N=10 (healthcare workers involved in implementation; medical officers/doctors n=5, microbiologists n=3, lab techs n=2)	to explore enablers and barriers in using Xpert among the targeted groups from the providers' perspective	interviews
McDowell	2018	India (lower middle icome)	Urban	Newly established high throughput low-complexity automated NAAT labs, one per city, which linked to public/private clinics and hospitals	Both public and private	High burden	Xpert MTB/RIF	study focuses on a project to improve the implementation of upfront Xpert testing for pediatrics by free testing with quick turn-around times (within 24hrs) and efforts in coordination with local authorities to improve provider literacy to diagnosing tb in kids	pediatric presumptive TB patients with fever more than 2 weeks, unremitting cough for more than 2 weeks, and/or weight loss or no weight gain in past 3 months	N=55 (physicians who had referred samples for Xpert testing (20 public, 22 private physicians, 5 trust hospitals, 8 TB programme officers)	To better understand the perspective of providers engaged under the ongoing project with respect to Xpert testing, related national and global guidance for the diagnosis of TB in children, and various bottlenecks in its effective implementation. i) how do pediatricians use Xpert when accessible and free of cost, ii) how do they prioritize and evaluate Xpert in relation to other diagnostic technologies, and iii) what are the effects of Xpert on their clinical practice	semi structured interviews
Shewade	2018	India (lower middle income)	Both rural and urban	low-complexity automated NAAT testing at tertiary district level facility; sputum smear at microscopy centre where samples need to be sent from	Public facility	High burden	low-complexity automated NAAT and LPA	In January-March 2014, if sample was smear positive, then LPA was used upfront. Among smear negative samples, culture was done followed by LPA. From April 2014 onwards, LPA was used for smear positive and CB-	presumptive MDR-TB patients/high risk patients	N=23 (lab technicians n=6, treatment supporters/supervisors n=12, microbiologist n=2, district TB officer n=1, senor Tb lab supervisor n=1, senior DR	To explore from the healthcare provider perspective, the barriers and suggested solutions for improving DST in programmatic setting in Bhopal district, India.	interviews (10), FGDs (2) plus one FGD to discuss solutions later
Joshi	2018	Nepal (lower middle income)	Both rural and urban	low-complexity automated NAAT centres in district hospitals, primary health centres, district public health office laboratory	Public facility	Unclear/not reported	Xpert MTB/RIF	not reported	children (<15 years); people living with HIV (PLHIV); severe forms of TB; and presumptive MDR TB patients	unclear: 22 interviews (Presumptive TB patients), 4 Focused group discussions (district TB officer and/or lab personnel in four centres) In two centres, the FGDs were replaced by IDIs due to small numbers of participants. National level-In depth interviews n=4 (NTP focal person for Xpert MTB/RIF, monitoring and evaluation section chief, WHO focal person for NTC and TB coordinator for International Organization of Migration)	To explore the barriers to effective implementation of the Xpert MTB/RIF assay (mixed methods sequential explanatory design a qualitative evaluation)	FGDs, in-depth interviews, semi structured interviews(patients)
Stime	2018	South Africa (upper middle income)	Urban	Busy public clinic	Public facility	High burden	Xpert MTB/RIF (16-module)	on site 16 module Xpert machine, batching samples in 2-3 runs per day (approx 48 samples/day) and in parallel HIV rapid and viral load testing (some STI testing ongoing, chlamydia and gonorrhoeae with Xpert as well)	Not reported	N=20 (clinic staff: nurses n=6 physicians n=2, laboratory technicians n=5 administrators n=5, security guards n=2	to describe clinic flow with special emphasis on the impact of POC testing at a large urban public healthcare clinic in Durban, South Africa. (mixed method, time in motion study)	semi-structured interviews

First author	Year of publication	Country (income classification)	Geographical setting	Type of health facility	public/private	Background prevalence	Diagnostic technology	Programmatic features of the intervention (Where and how)	Target population	Total number of participants and types of settings	Research questions/ objectives	Data collection methods
Rendell	2017	Mongolia (lower middle income)	Urban	National TB reference lab, provincial TB clinics, district TB clinics, hospital	Public facility	High burden	Xpert MTB/RIF	unclear. only eligibility criteria are reported and a weekly consensus meeting for treatment initiation is mentioned	All smear negative pulmonary TB patients; Patients with presumed pulmonary TB diagnosed with HIV/AIDS; patients with presumed MDR- TB or XDR-TB; (All smear positive new patients aged 15-34 years old (this guideline is yet to be implemented))	N=24 (laboratory staff n= 8, TB physicians n= 16)	to identify and understand system and context specific factors within Mongolia's National Tuberculosis Program (NTP) that are barriers or enablers to implementing the Xpert MTB/RIF test from the perspective of NTP staff.	semi-structured interviews
McDowell	2016	India (lower middle income)	Urban	Clinic	Private facility	High burden	Xpert MTB/RIF, sputum smear, Xray	highly variable, mostly empirical treatment first, then a range of tests, always including Xray. If low-complexity automated NAAT then late in the process	Patients with presumed pulmonary TB diagnosed with HIV/AIDS	N= 185 (private providers- different specialization n-110, patients n=75)	To understand the factors contributing to the variability in care and the presence of practices diverging from the standard of TB care in India.	interviews, observations of clinical practice and Continuing Medical Education events
Engel	2015	South Africa (upper middle income)	Both rural and urban	Clinics and hospitals	Both public and private	High burden	Xpert MTB/RIF at POC in district hospitals	Patients present at different levels of care (in clinics, health posts, laboratories or hospitals) with multiple or unspecific symptoms (e.g. acute febrile illness) and may need several diagnostic tests. Testing mostly centralised but in some district hospitals Xpert is available.	Patients with presumed TB	N= 141 participants ({interviews n=101 with doctors, nurses, community health workers, patients, laboratory technicians, policymakers, hospital managers and diagnostic manufacturers} and {focus groups n =40 with TB patients, nurses and community health workers}, interviews not focused on TB diagnostics exclusively	To examine POC testing across major diseases in South Africa contributing to burden of disease (mainly HIV, TB, diabetes mellitus, diarrhoeal diseases and hypertension). We assessed what tests are performed and how in public/private, rural/urban hospitals and clinics and whether they can ensure successful POC testing.	Interviews (101), Focus group discussions (40)
de Camargo Jr	2015	Brazil (upper middle income)	Urban	Clinic	Public facility	High burden	Xpert MTB/RIF	not reported	presumptive TB patients	unclear; In Rio de Janeiro, interviews with 11 patients diagnosed with smears and 9 diagnosed with Xpert MTB/Rif. In Manaus, 10 interviews with patients diagnosed with Xpert. In Rio de Janeiro, a physician, a nurse, a laboratory technician and an administrative staff member were involved in the preparation of the flowcharts. In Manaus, the director of the facility, one municipal health official, a physician, two nurses, a receptionist and a lab technician participated. 3 field researchers participated in all group meetings to elaborate the flowcharts. They also interviewed key informants at the research sites and higher-ranking positions of local health departments (number not specified).	To qualitatively evaluate the repercussions of the adoption of the Xpert MTB/Rif in the Brazilian Health System from the perspective of patients, health professionals and managers, considering aspects such as understanding, perception and meaning	Interviews, group meets to produce diagnostic flowcharts
Naidoo	2015	South Africa (upper middle income)	Urban	Primary health-care facilities, (central laboratory)	Unclear/not reported	High burden	Xpert MTB/RIF, LPA	the testing algorithm changed during the study: In 2010, a smear, culture and LPA-based diagnostic algorithm was used with LPA done on culture isolates or clinical specimens of high MDR-TB-risk presumptive cases (those with previous TB, an MDR-TB contact or from a congregate setting). From 2011–2013, Xpert was phased in, replacing smear microscopy for all presumptive TB cases	presumptive TB patients	N=26 patients	to explore and compare MDR-TB patients' experiences of their pathway to diagnosis and treatment initiation in LPA and Xpert-based diagnostic algorithms.	Interviews
Hoang	2015	Vietnam (lower middle income)	Unclear/not reported	District health centre/hospital	Public facility	High burden	Xpert MTB/RIF	unclear	Patients at high risk for MDR-TB	N= 110 (TB provincial staff members n=30 health staff N=80 [8 central, 56 provincial,16 district & community level])	To understand challenges of efficient implementation of 5 steps from diagnosis to MDR-TB treatment (mixed method study)	Focus group discussion, interviews, document review

First author	Year of publication	Country (income classification)	Geographical setting	Type of health facility	public/private	Background prevalence	Diagnostic technology	Programmatic features of the intervention (Where and how)	Target population	Total number of participants and types of settings	Research questions/ objectives	Data collection methods
Colvin	2015	South Africa (upper middle income)	Urban	Hospital and clinic	Public facility	High burden	Xpert MTB/RIF and GenoType LPA (HAIN)	upfront TB testing with Xpert, but ultimately located primarily in laboratories and not primary care clinics	presumptive TB patients	N=40 informants (Global N=3{WHO N=1, FND N=1, Cepheid N=1 }, National level N=3, Provincial level N=2, District N=16 {sub district manager 3, TB programme manager N=1, TB hospital manager N=2, TB-HIV sub-district coordinators N=4, DR TB nurses N=3, National health laboratory service N=2, MSF N=1} Health facilities N=16 { facility manager =4, TB/DR TB doctors N=5, Nurses N=3,TB clerks/assistants N=4})	To examine policy transfer for GenoType LPA and Xpert to understand how these promising new technologies were taken up, adapted and delivered within local health systems	longitudinal, qualitative evaluation to track policy transfer with the introduction of Xpert and GenoType LPA in South Africa. Two phases of key informant interviews were complemented with reviews of quarterly reports from health and laboratory services and other relevant documents.
Royce	2014	Cambodia (lower middle income)	Unclear/not reported	Regional laboratories, district and referral hospital and health centres	Unclear/not reported	High burden	Xpert MTB/RIF	Cambodia's guidelines recommend that previously treated patients have sputum specimens tested using Xpert MTB/RIF (available in four provincial laboratories), followed by culture and species identification using liquid and solid media (available in three regional laboratories) and conventional DST at the national reference laboratory.	Previously treated TB patients	Unclear; interviews N=26 (doctors or clinical officers n=9, nurses n=8, laboratory staff n=6, and TB officers n=3). Focused group discussions (N= unclear)	To quantify the gaps in the detection of MDR-TB in previously treated TB patients in Cambodia, and describe health workers' perspectives on barriers, facilitators and potential interventions sequential explanatory mixed-methods design	FGDs and interviews
Creswell	2014	Nine countries (Democratic Republic of Congo (DRC)-low income, Kenya- Lower middle income, Pakistan- lower middle income, Bangladesh-lower middle income), Mozambique-low income, Cambodia-lower middle income, Malawi-low income, Nepal-lower middle income, Moldova-lower middle income)	Both rural and urban	District hospitals, laboratories, AIDS centres,	Both public and private	High burden	Xpert MTB/RIF	different approaches per country (active, passive, mixed, screening); Placements included public and private hospitals and lower primary care facilities, private diagnostic laboratories, HIV centres, prisons, reference laboratories and mobile units. The projects were able to run the machines at district hospitals and at lower levels of care although in only a few situations were peripheral microscopy centres included, mostly because of throughput concerns.	MTB/RIF testing on patients with suspected TB, implementation variable, mostly sputum negative	unclear; project staff, implementers, manufacturer	To present results from nine TB REACH interventions, review the main challenges experienced and formulate recommendations for other early implementers mixed methods,	Document review and semi structured interviews with staff from each project and manufacturers.
Jaroslawski	2012	India (lower middle icome)	Urban	Peripheral laboratory, clinic	Both public and private	High burden	Serology tests, molecular tests such as Xpert MTB/RIF	unclear, private providers use serology, Xpert is mentioned as one of the costly alternatives	presumptive TB patients approaching private providers	N=41 (private doctors and private hospital laboratory staff (n= 11), private stand- alone laboratories (n= 7), distributors of diagnostic tests (n= 7), manufacturers of diagnostic tests (n= 7), government hospital doc-tors (n= 4), and NGOs working in TB (n= 5)	To explore why serological tests are so popular in the private sector and what factors have paved the way for their widespread use	interviews

Web Annex D.11. User perspectives on nucleic acid amplification tests for tuberculosis and tuberculosis drug resistance: (interviews study)

Authors: Ricky Janssen, Muthoni Mwaura & Nora Engel
Department of Health, Ethics and Society, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, Netherlands (Kingdom of the)

Produced in preparation for the WHO guideline group meeting on "Nucleic acid amplification tests to detect tuberculosis and drug-resistant tuberculosis" on 7-18 December, 2020.
Please send comments and feedback to Nora Engel, n.engel@maastrichtuniversity.nl

Introduction

A fundamental component of addressing the rising burden of drug-resistant tuberculosis (DR-TB) is expanding the coverage of drug susceptibility testing (DST) to all persons with signs and symptoms of TB (WHO, 2020). In doing so, it is not enough to only assess the accuracy of new diagnostic technologies but it is vital to consider user perspectives on the value, feasibility, usability, and acceptability of these technologies. If the perspectives of laboratory personnel, clinicians, patients, and TB programme personnel are not considered, these technologies risk being inaccessible to and underutilized by those for whom they are intended.

Qualitative research provides a useful platform to explore the user perspectives of key TB diagnostic stakeholders. Using targeted sampling, qualitative research is able to assess the diagnostic values and experiences of various users in multiple settings. It is well understood that the value of new diagnostics is not inherent within the technologies themselves, but rather through the alignment of technology with the specific needs of users in particular settings. Qualitative methodology provides a window through which user experiences can be captured, analyzed, and translated into workable data for policymakers to consider. By engaging users through interviews, ethnographies, usability tests and other methodology, qualitative studies can support decision-making on diagnostics and offer concrete insights into users' values and preferences, as well as acceptability and feasibility of new diagnostics in the intended use setting.

In December 2020, the World Health Organization assessed three classes of nucleic acid amplification technologies to detect TB and DR-TB. To inform those discussions, the WHO commissioned a study into the perspectives, preferences, and experiences of users of low-complexity automated nucleic acid amplification tests (NAATs) for detection of resistance to isoniazid and second-line anti-TB agents, medium-complexity automated NAATs for detection of TB and resistance to rifampicin and isoniazid, and high complexity hybridization-based NAATs for detection of resistance to pyrazinamide (PZA LPA). To this end, we conducted a small qualitative study with participants in India, Republic of Moldova, and South Africa. We interviewed clinicians, laboratory staff, programme officers, and MDR/XDR TB survivors with the aim of understanding their experiences of using these various technologies as well as their general TB diagnostic experiences.

Methods

During October and November 2020, RJ and MW conducted 14 semi-structured interviews with clinicians, programme officers, laboratory staff, and patient advocates in India, Moldova and South Africa. These countries were selected based on the fact that they fall on WHO's list of 30 high MDR-TB burden countries (StopTB Partnership, 2020) and the index tests have been used to some extent in research contexts within these three countries. Due to the short time-frame, participants were purposively sampled and approached based on convenience through personal contacts and colleagues. For an overview of the participants, please see Table 1. Professional roles are used to mask study participants' identity, coded by their country (Moldova (M), India (I), or South Africa (S)), their profession (clinician or medical doctor (M), patient advocate/representative (R), laboratory staff (L), or programme officers (P)), and a number.

Interviews were conducted over Zoom, Skype and phone. Topics discussed included: 1. current approach to diagnosing TB, MDR-TB, and XDR-TB including specific challenges; 2. experiences with using molecular TB diagnostics and the index tests specifically, including details on steps taken in the diagnostic process; 3. determining eligibility and treatment initiation as well as challenges and benefits of using the index tests; 4. overall usefulness of the index tests; 5. the feasibility of implementing the index tests; 6. their potential impact on health equity and; 7. how this relates to current policy context.
We note several important limitations of this approach. We were only able to interview a limited number of participants per setting and country. Owing to the use of Zoom, Skype, or phone for interviews, we were not able to triangulate interview data with other evidence commonly collected through ethnographic approaches (such as multiple interviews and informal conversations at the same facility, observations or site visits). In addition, not all participants had personal experience with one or all of the index tests, and those participants who did have experience with the tests had used them in research settings and not for routine practice.
Interviews were audio-recorded, transcribed by MW, and coded by RJ in NVivo. The authors each wrote memos on different topics, discussed these and collated them into themes, which we present in the results section below.
Table 1: Participant Overview

	Moldova (M)	India (I)	South Africa (S)
Clinicians (M)	1	1	1
TB survivor/Advocates (R)	1	1	1
Lab personnel (L)	2*	5*	2
Programme officer (P)	2*	2	1

* Were interviewed as a group

Ethics
This study was approved by FHML-REC, the ethical review board of the Faculty of Health, Medicine and Life Sciences at Maastricht University (FHML-REC/2020/105). Study participants were emailed an information sheet explaining the objectives of the study and an informed consent form which they were asked to sign and return prior to participation.

Results

The results section reveals findings from our interviews with laboratory staff, clinicians, programme officers, and patient representatives from India, South Africa and Moldova. The first three sections present results for a given index test, where we elaborate on benefits and challenges of each test, as well as the feasibility and acceptability of each technology. However, results regarding these index tests need to be considered in relation to other parts of TB testing and care such as sample collection and transport, communication of diagnostic results from the laboratory to clinicians, and the potential impact of these index tests on health equity. Therefore, in section four, we consider results regarding sample collection, transport and quality. In section five, we discuss how laboratory staff and clinicians make sense of diagnostic results, and finally, in section six, we reflect on the overall impact of diagnostic changes for TB in relation to health equity and patient care.

Part 1: The Index Tests

Index Test 1: Low-complexity automated NAAT

The first class of technologies to be evaluated at the WHO 2020 guideline meeting are low-complexity automated NAATs for the detection of resistance to isoniazid (INH) and second-line anti-TB medicines. One such test that we investigated in this study is the Xpert MTB/XDR cartridge by Cepheid. This particular cartridge detects mutations associated with resistance towards INH, fluoroquinolones (FLQ), second-line injectable drug (SLID) (amikacin, kanamycin, capreomycin) and ethionamide (ETH) (Cepheid, 2020).

Benefits

Although only three participants in this study had used the cartridge in a research setting (ML-1, ML-2, SL-2), almost all the participants highly anticipated its use for routine care. When asked what are some anticipated benefits of the cartridge, the most frequent response was its faster turnaround time (TAT) for a drug susceptibility test (DST) result when compared to existing molecular and phenotypic technologies (i.e. first and second-line line probe assay [LPA], whole genome sequencing, solid culture, and liquid culture; IL-1, IL-2, IR-1). Participants further linked this faster TAT to the possibility of initiating the appropriate treatment for MDR and XDR patients sooner than has typically been the case (IL-4, IM-1, IP-1 IR-1, MP-2, MR-2, SM-1, ML-1, SL-2), a value that was especially anticipated by the XDR TB survivors who participated in the study (IR-1, MR-2).

> *"I think that it's absolutely crucial to have the proper, a quick diagnosis, the results of a diagnosis should be received as quick as possible as to inform a proper regimen…As a patient I can't, for example afford myself to take pills for three or four months just in vain and being informed about the right type of resistance only in a couple of months because, maybe generally it is not much time and for clinicians two or three months doesn't make a [difference], but for patients taking a handful of pills daily for two or three months just in vain, that's not fair to take, if they are not helping"* (MR-2).

Laboratory personnel highly valued the minimal user steps, minimal technical training and expertise, and minimal laboratory infrastructure that Xpert MTB/XDR needs in comparison to LPA (IL-2, ML-2, IL-4). The Xpert MTB/XDR cartridge seemed familiar to them because of its similarities to the existing Xpert MTB/RIF and/or Xpert MTB/RIF Ultra. This therefore meant that it would be easy to run and the results produced by the platform would be straightforward to interpret (IL-2, ML-2). They anticipated that this would also allow for the cartridge to be positioned lower in the laboratory network than culture and LPA, allowing for the decentralization of drug resistance detection, treatment and monitoring (ML-1, IM-1,). Furthermore, a laboratory technician said that the modular-based structure of the platform gives it the flexibility to be scaled-up or down, depending on the needs of the TB programme and laboratory network (SL-2).

Both laboratory personnel and clinicians found the kind of information that the diagnostic provides to be valuable, particularly the level of resistance to the anti-TB drugs (i.e. low resistance or high resistance; ML-2, SL-2, and SM-1). The number of drugs that are tested on the machine and the types of mutations detected were also important, particularly in regards to INH resistance. One laboratory scientist who had experience with the Xpert MTB/XDR cartridge noted that it had an "edge over other molecular technologies" because it assessed multiple genetic targets associated with INH resistance (SL-2), and a clinician in India with no experience with the cartridge anticipated that it could help in the detection of INH monoresistance, which he reported to be a burden in India (IM-2). Knowing the resistance of fluoroquinolones and ethionamide early

was also important in giving clinicians necessary information to develop appropriate treatment regimens as soon as possible (SM-1, IM-2). A laboratory scientist (SL-2) summarized the value of Xpert MTB/XDR as:

> *"you would get almost universal resistance prediction for a resistant patient, within 6 hours. Because you would have RIF, you would have INH, you would have fluoroquinolones, you would have aminoglycosides, you would have partially predicted ethionamide, clinicians have got a lot that they can work with, from a primary sample from a patient".*

Challenges

When it came to experienced and/or anticipated challenges of the Xpert MTB/XDR, participants most often reported the cost (IL-1, IP-2). Because Xpert MTB/XDR cartridge runs on the 10-colour module, programmes would need to replace and/or compliment their existing GeneXpert platforms, a consideration that bears significant financial implications (SL-2). Because of the cost that this would bring, as well as the relative lower burden of MDR/XDR TB in comparison to MTB, most participants agreed that the Xpert MTB/XDR cartridge would not be feasible if positioned as a baseline MTB test, but rather at the district level or intermediate reference laboratory level where current DST takes place (IL-1, IP-1, IP-2, IM-1, MP-2). Therefore, despite the cartridge's easy user steps and few infrastructure requirements allowing for a decentralized positioning, participants cautioned that if placed too low or peripheral in the network, the test risks being underutilized, a finding similar to that echoed by quantitative and qualitative studies on the utilization of Xpert MTB/RIF (Cazabon et al., 2018, Mwaura et al., 2020, Albert et al., 2016).

Instead, a programme officer in Moldova proposed an algorithm that positions the cartridge higher up in the laboratory network where it can receive samples that test 'MTB detected, RIF resistance' on the Xpert MTB/RIF machines located more peripherally (MP-2). Not only may this require the patient to provide an additional sample, but it also would require an efficient sample transportation system, both of which are challenges noted by participants in this study (IL-1, SM-1, SP-1, SL-1). This is reflected in Findings 9, 10, and 17 in the report by Engel et al. (2020), where they note that challenges related to obtaining a second sample from a patient as well as poor/non-functioning sample transport mechanisms result in diagnostic delays and underutilization of low-complexity automated NAATs. Additionally, placing the Xpert MTB/XDR cartridge at a different location from the Xpert MTB/RIF cartridge does not allow the laboratory technician to reflex the sample to the XDR cartridge. A laboratory participant mentioned reflexing the sample as a potential benefit, if the system is set-up to report the result directly to the laboratory personnel (SL-2). Nonetheless, participants of this study value the anticipated benefits of the Xpert MTB/XDR test as a molecular DST, and eagerly await its introduction into routine diagnostic care.

Low-complexity automated NAAT Summary: Acceptability and Feasibility

The Xpert MTB/XDR cartridge addresses several preferences/values of laboratory staff and clinicians. It requires minimal user steps and the GeneXpert platform is a familiar system which people feel comfortable running and interpreting. The cartridge has a quicker turnaround time for first and second line drug susceptibility testing, compared to other available diagnostic methods. People value faster TAT, the potential ability to reflex samples from the Xpert MTB/RIF to the Xpert MTB/XDR cartridge, and receiving information on multiple drugs as well as high or low level resistance simultaneously, as it could enable quicker diagnosis and optimize treatment for patients. The Xpert MTB/XDR cartridge appears widely acceptable by laboratory

staff and clinicians based on its simple user steps, and due to the amount of important information it provides. The new cartridge requires less user training compared to other DST methods (such as LPA and culture), making it more feasible to implement. However, the cost of the 10-colour system for this cartridge, and the value of diagnosing MTB over DR TB at primary care, makes it less feasible as a baseline test, and therefore may be more suitable if placed at a district or intermediate level lab.

Index Test 2: Medium-complexity automated NAATs

The second class of technologies to be evaluated at the WHO 2020 guideline meeting are medium-complexity automated NAATs for the detection of TB, as well as resistance to rifampicin (RIF) and isoniazid (INH). These platforms are characterized by automated DNA extraction, PCR preparation, and result interpretation, and can run a minimum of 24 samples at a time. Examples of these platforms, which we discuss further below, include BD MAX MDR-TB (BD), Cobas MTB-RIF/INH (Roche), and FluoroType MTBDR Version 2.0 (Hain).

Benefits

Only one laboratory technician in South Africa spoke directly about their experience using medium-complexity automated NAATs (SL-2), while two other participants commented on the potential impact of specific medium-complexity automated NAATs (IP-2, SL-1). Laboratory staff valued medium-complexity automated NAATs due to their automation. One participant illustrated this while talking about the FluoroType platform. The closed-system platform would reduce "*all this moving around*" and therefore minimize labour for laboratory staff, which would be a "*great improvement over the current LPA*" (SL-1). Another participant described the FluoroType system saying;

> *"[…] There's the preprocessing unit which takes care of DNA extraction as well as loading your plate. So it adds the reagents for the PCR as well as the DNA. So it's neat right, it's basically a robot. So all you need to do is load your sputum samples together with the buffer, and the rest, you close the instrument, and the rest is taken care of." (SL-2*)

Platforms are valued for their simplicity and for their ability to fit into the laboratory setting. For example, a participant described the BD MAX platform saying:

> *"so the BD MAX, nice sized platform. Sits well on the bench. Up to 24 isolates per run. What's interesting is the preprocessing is about 20 minutes longer than your Cepheid, similar type of methodology, so pre-prep is as good as Cepheid. Then loading the cartridge, relatively easy. Setting up the instrument, similar, it's quite intuitive, good interface, really user-friendly." (SL-2)*

The BD MAX also produces a report which, not only provides an indication of TB detection or drug susceptibility, but also contains melt-curves, which can be used to predict resistance (SL-2). However, it is questionable whether laboratory staff would generally have the expertise or time to read melt-curve reports, and therefore the usability of this information in a routine laboratory setting is not clear. The report also contains information that allows the laboratory technician to infer the level resistance (e.g. low or high level resistance). The BD MAX design acknowledges the busy nature of laboratory work in two ways. First, its automation frees up time by allowing staff to walk away and work on other things in the meantime (SL-2). Secondly, the platform includes external lights (blue and red) which indicate to laboratory staff if everything is okay (blue) or if something is wrong (red) (SL-2). Therefore, the platform uses lights to get the attention of staff even if they are busy doing something else.

The flexibility of medium-complexity automated NAATs is a potential benefit. For example, one program officer from India mentioned that the COBAS platform had already been used for COVID-19 testing. He mentioned the possibility for different programmes to share the machine, but also share the responsibility of cost. For example, the TB programme could be primarily responsible for "consumables" such as the buffers, controls etc. related to TB testing (IP-2). In addition, he mentioned that medium-complexity automated NAATs could free up low-complexity automated NAATs for areas where they are currently lacking and ensure increased access to diagnostics.

Challenges

Potential challenges around medium-complexity automated NAATs include the cost of acquiring the machines and provision of good quality samples (ML-1, IL-1, IL-2). Furthermore, the chosen platform has to align with the physical laboratory space. For example, although the Fluorotype platform was valued for its automation, a participant mentioned concerns regarding the size of the platform.

> *"So you just can't place it anywhere. You can't just place it on any bench top. But fortunately I work in a very nice laboratory, it was funded by USAID, PEPFAR, so I have got a world-class smart, you know the doors are well designed to allow huge large instrument pieces into the facility, things like that… I have double doors and things. In our routine environment, I don't see it easy to install this type of instrument, if the doors are standard frame. They're massive." (SL-2)*

The platform was described as having a "massive foot print". It was too large and heavy to sit on a standard lab-bench or fit through a door frame. The thermocycler component of the system is very large. Part of the reason this element ended up being so large is that it was designed to withstand high heat and dust.

> *"[…] so the idea was that it could run in any environment in any country, particularly a country with harsh temperatures and dust and stuff like that. So it was a robust, like in the field type of use. Which I don't see people installing it under a tree in Africa, despite it being built for that type of purpose, but yeah, I think labs generally across the world are not that bad where you actually need to take so much care." (SL-2)*

This illustrates the tension in the design of the machine: for higher-level reference laboratories dust and heat are not likely to be an issue. The size of the machine in this example outweighs the benefits of its more robust environmental operating capacity, as this is not necessarily a concern for high-level laboratories with good infrastructure.

Another consideration relates to where these medium-complexity automated NAATs should be placed within the larger laboratory infrastructure. A system like the BD MAX for example is not modular, and once the run has started one cannot add additional samples. When talking about the BD MAX system, one participant stated that if you did not have a minimum of 12 samples to run, you would be wasting reagents. TB programmes therefore need to consider placement of these medium-complexity automated NAATs in laboratories central enough that they receive adequate sample numbers to make the machine worth running. Yet, this requires a well-functioning sample transport system.

Medium-complexity automated NAATs Summary: Acceptability and Feasibility

Medium-complexity automated NAATs address several preferences/values of clinicians and laboratory staff; it is faster than culture DST (like LPA or cartridge-based tests); has the advantage of being automated (unlike LPA); and gives additional clinically-relevant DR information e.g. high vs. low resistance (unlike the current GeneXpert MTB/RIF cartridge) (also see section V.). Automation, which reduces staff workload, the ability to run multiple samples at once, and the detailed report information provided by medium-complexity automated NAATs, contribute to the acceptability of these platforms within a high-level laboratory. For programme managers/officers, one additional benefit of medium-complexity automated NAATs relates to flexibility of use and potential use across multiple disease programmes. However, the physical size of the platform and how it fits in the laboratory space affect this acceptability (smaller footprint is better), and can affect the feasibility of fitting this kind of platform into a laboratory space. A functioning sample network also challenges feasibility of implementing medium-complexity automated NAATs and laboratory technicians voiced concerns over sample quality (please also see section VI. and findings 9 and 10 in the report by Engel et al., 2020). Additional feasibility considerations for this method include clear communication of diagnostic results for clinicians (also see section V.) and ensuring the laboratory where it is placed is central enough to receive adequate numbers of samples to make the machine worth running.

Index Test 3: High-complexity Hybridization-based NAAT (PZA LPA)

The third index test being evaluated at the WHO 2020 guideline meeting is high-complexity hybridization-based NAATs for the detection of pyrazinamide resistance (PZA LPA). An example of this test is the Genoscholar PZA-TB II assay (NIPRO).

Benefits of PZA LPA

Only one participant (SL-2) had experience using the NIPRO test, and this was within a research context in South Africa. However, laboratory participants did have experience doing line probe assays for first and second line DST, and were therefore able to comment on potential benefits and challenges of PZA LPA. Potential benefits of PZA LPA identified include lower cost compared to whole genome sequencing; easier interpretation than whole genome sequencing; faster than culture DST; and the possibility of providing information on PZA resistance at more than one concentration.

> *"[…] we do not have a currently reliable technology for PZA, and PZA being one of the important drugs both in clinical first line treatment or second line treatment. I think it's good to have something which can give us fast results for PZA resistance. So that way LPA it can be advantageous provided that we can develop an LPA technology that is fairly predicts the resistance." (IL-2)*

Aside from high-end sequencing, which is too costly for settings such as Moldova (ML-2), culture DST is currently the only other method of determining PZA resistance, and this can take weeks. LPA methods only take 1-3 days, depending on the laboratory. Therefore, an LPA would significantly cut down on turnaround-time for PZA DST. In addition, current culture methods commonly being used for PZA resistance testing include BACTEC MGIT. This only provides information regarding PZA resistance at one concentration level (ML-1). One participant from a laboratory in Moldova also highlighted the benefit of having a quicker resistance test for PZA even though it is not a new drug, as new TB drugs are "not developed everyday" (ML-2).

Specifically in relation to the PZA LPA from NIPRO, one laboratory participant who used the system in a research context in South Africa mentioned that they appreciated the small footprint of the design (SL-2).

> *"[…] the nice thing about the Japanese version of the line probe assay is that the footprint is much smaller. […] the line probe assay on NIPRO, much tinier. Size of a microwave. Whereas the GT-Blot is two microwaves, yeah about 40-liter microwaves, two of them. This is your NIPRO single, less complicated-looking system, and works really well." (SL-2)*

Similar to the medium-complexity automated NAATs, having a platform size that fits well within available laboratory space is important.

Challenges regarding LPA and how this relates to PZA LPA

Although LPA in general is a faster method for prediction of first and second line drug resistance than culture DST (1-3 days vs. up to 6 weeks), participants raised several issues regarding this method. First, the LPA method requires significant training and expertise from laboratory staff. It also requires significant laboratory infrastructure including multiple rooms to reduce the risk of sample contamination, and requires good quality samples. The ability of LPA to accurately detect drug resistance in smear-negative samples is questionable, which has led to instances where LPA is only run on cultured samples, thus undermining the benefit of LPA's faster turnaround time. Finally, interpretation of LPA is complex, requires significant training, and can still pose issues due to lack of confidence in results.

> *"I mean the line probe assay at its time was great. You know, it was better than anything available, or there was nothing available at the time. But of course the newer technologies are definitely going to replace it, most definitely." (SL-2)*

For LPA to be feasible, it requires good laboratory infrastructure, with separate spaces to mitigate the risk of contamination, as well as highly trained laboratory personnel.

> "*But if you take LPA type of process, I think that has a little more higher technical skills, because like [the assay? 42:38] isolation of DNA first and then you perform a PCR and then you do a hybridization step, so there are three steps involved where things can you know, it requires safe hands you know, experienced hands to do the assay. […] Because many times the patient samples have [inhibitors? 44:10], so if it is not handled well, that can give [very interesting? 44:15] inconclusive results and I think I don't know how common it is in LPA, but in our past experiences in LPA, we have seen a lot of inconclusive results and we have to go back and re-start our experiment and repeat it. So it is very critical if there is even dust from the environment which is coming. So you prefer slightly good working conditions for that kind of an assay." (IL-2)*

Sample Quality

For an LPA to provide an accurate result, skilled technicians need to collect a good quality sample, transport it in time and then handle the sample in a way that avoids contamination, which may lead to inconclusive

results (IP-2, IL-2) (also see Finding 10 in report by Engel et al., 2020). Proper sample handling, as well as proper temperature maintenance, and use of correct volumes of liquid are crucial in order to obtain accurate results (IL-2). LPAs also require that part of the sample is dedicated to controls, and that there is sufficient sample volume to make it worth running controls from a cost perspective (IL-1). The lower the sample load, the more controls need to be run, resulting in a higher overall cost.

Furthermore, only samples with a high enough DNA load can be run using the LPA, further limiting its application based on the sample quality, which means the LPA is only run on microscopy positive (one plus) samples to ensure validity of the test (ML-2). Some smaller laboratories in South Africa only run LPAs on cultured samples, due to the high number of false positives from direct testing (SL-1). One participant suggested that this relates to sample contamination and their inability to get rid of the contaminants (SL-1). In larger laboratories, the LPA is run directly on the sample depending on the microscopy and Xpert result. One of the main considerations around whether a laboratory is able to do direct LPA on a sample relates to how well the laboratory can mitigate contamination and concerns regarding the usability of LPA on smear-negative samples (SL-1).

> *"yes, this is personal point of view, from experience as a lab, I am always nervous of direct LPA on smear negatives. I hope they prove me wrong on the new technologies, even when we are doing it, I'm always skeptical on smear negatives with my experience…" (SL-1)*

Difficulties Interpreting Results

Furthermore, test interpretation of LPA is made more difficult due to the issue of interpreting resistance when wild-type is not present, yet lack of mutations suggest drug susceptibility. This is made more complicated when read in the context of a discordant Xpert and/or phenotypic result.

> *"[…] for example we have this case, if it's for example at LPA wild-type three and four are missing, and no mutation is present, usually according Hain its resistance, but to the GeneXpert sometimes and also to the phenotypic, its sensitive. So we have like discrepancies between the testing." (ML-2)*

Dealing with these kinds of discrepancies takes work between the laboratory staff and clinician (also see section V). This work includes monitoring patient progress on a certain drug regimen, taking additional samples, and running new tests. A South African clinician illustrates this below saying;

> *"[…] so the first sample that was sent, there was a discrepancy between the LPA, it said that he was resistant, and I think the liquid culture said that he was sensitive. So we weren't sure which one to believe. But the patient I think had been on treatment for maybe three months, and he was doing clinically well, but because of that result, you know the clinician decided to change him to the longer regimen, and then send off a second sample, and the second sample showed that both the LPA and the culture were sensitive, and then the matter was discussed with the infectious disease specialist, the microbiologist at the lab, to try and decide what was the way forward. And then the plan was that there was contamination with the first sample." (SM-1)*

This example also illustrates how discordant results can lead to multiple treatment regimen changes, which can be confusing for the patient. Laboratory staff have also experienced reliability issues with the results provided by the HAIN LPA. For example, there were times when the report suggested drug susceptibility through the production of a light band, where in reality they found that these light bands were indicating rifampicin resistance (SL-2).

For one participant who had direct experience using the NIPRO PZA LPA, he mentioned concerns regarding how to interpret the result of the test.

> *"the PZA for NIPRO, there was just too many bands it's ridiculous. You know, because the pncA mutation span across the entire gene land, there's multiple mutations related, it's like no hotspot. And what happens is like actually interpreting those line probes strips is ridiculous. It's too many probes on one strip, compacted and for you to make proper, it would take me about 10 minutes or five minutes because I am an expert to actually interpret the banding pattern on those strips. You know, too many targets." (SL-2)*

The Hain LPA comes with an AutoReader that helps the user interpret the results of the LPA. The NIPRO PZA LPA did not come with an AutoReader to predict resistance and susceptibility. NIPRO did develop a mobile phone app to interpret the test, but the development and implementation status of the app is unclear (SL-2). The participant added that an app for interpretation would fit well into a lab setting.

Testing for PZA Resistance: Clinical Relevance and Routine Use

The relevance of PZA resistance testing in general seemed to vary between countries. In South Africa, participants often mentioned that PZA resistance testing is only done for surveillance and research, but is not routinely utilized for clinical settings (SL-2, SM-1, SP-1).

> *"nationally we do it as a reference lab for research projects particularly and also for surveillance activities but generally in the routine diagnostic practice we do not test for pyrazinamide, because irrespective of its resistance, it's still used." (SL-2)*

A South African clinician said that she would need to specifically request PZA resistance testing and that it was not routinely done, also due to cost (SM-1). Other reasons mentioned regarding lack of routine PZA DST include lack of confidence in PZA DST results, and possibly the high load of testing currently going on in laboratories (IP-1, MM-1). There were also questions regarding the justification for prioritizing PZA DST given the use of other, newer TB drugs.

> *"And I think it makes sense to then find out for all these patients, who are resistant to like our core drugs, like bedaquilline, linezolid, clofazimine…I would focus on those you know, than prioritize pyrazinamide, you know. I mean there are other things, like for example people use cycloserine in longer regimens, they use terizidone and nobody has ever tested these routinely. At this point in time testing clofazimine is not done routinely. So we should focus on tests DST for clofazimine, cycloserine, terizidone, you know, those drugs, bedaquilline, linezolid. I think we should focus on those rather than PZA" (SP-1)*

The participant mentioned that although we do not currently have a lot of information regarding the general incidence of PZA resistance, PZA is also no longer considered a core drug for rifampicin-resistant TB (SP-1). Meanwhile, a clinician/participant in India stated that in one study from 2015, up to 27% of sample isolates were found to be PZA resistant. He considered PZA resistance to be a problem, suggested that there is currently demand in the medical community for a molecular test on PZA, and mentioned that a faster PZA resistance test would be helpful in designing an appropriate and effective treatment regimen for his patients.

> *"[…] we have been asking basically for as you know there are issues regarding molecular testing to pyrazinamide as of now. We don't have molecular testing to pyrazinamide available to us at all. And only the culture DST is done. DST is done longer and even right now, in fact, the DST has been started in every lab now, but whatever the results we are getting for DST is by culture. And obviously the issue is they are after two months, two and a half months or three months you get the report that the patient, pyrazinamide is resistant. […] So, yes the huge problem is the absence of molecular testing for resistance to pyrazinamide." (IM-1)*

Implementing a new, faster test for PZA resistance requires that countries decide where this test will happen in the testing algorithm. One participant stated that the PZA LPA should be conducted at the same time in the testing algorithm as first-line LPA (IM-1). His reasoning for this was based on the fact that his treatment decision in relation to other first line drugs would change if PZA resistance was indicated and he would extend treatment to nine months (instead of six months), based on WHO guidelines. This comment therefore also reveals how PZA resistance testing early on in care can help the clinician make important decisions regarding how to develop an individualized treatment regimen.

Another participant mentioned that in India, LPA tests are not run at the district level, and that PZA LPA would be run at intermediate references laboratories, like the other LPAs (IM-1), while another participant said that there are 20 laboratories in India currently doing PZA DST where the PZA LPA could be implemented. Overall, there seemed to be a wide variety of perspectives regarding PZA DST. However, these results do suggest that if PZA LPA is implemented there needs to be clear communication regarding how it will fit into the testing algorithm. This reflects Finding 16 in the report by Engel and colleagues (2020), where they highlight the importance of inclusive and transparent decision-making processes to reach consensus on the strategic and equitable implementation of new diagnostics.

High-complexity Hybridization-based NAATs (PZA LPA) Summary: Acceptability and Feasibility

The PZA LPA addresses some preferences/values of laboratory staff and clinicians. It provides quicker results regarding PZA resistance, compared to other available methods (e.g. culture DST), it can provide information on different concentration levels, and targets a drug that is widely used in first-line TB treatment. However, clinicians, programme staff and laboratory staff raised several concerns regarding acceptability and feasibility of this method and drug target. Acceptability of this method is dependent on; how well it performs on different samples, as laboratory staff question how well LPA methods work on smear-negative samples; how well it

actually detects mutations specific to PZA resistance; and clarification in some settings as to why this specific DST drug test is being prioritized, as it is not currently part of routine DST. In regards to feasibility, participant concerns centered around the significant training and laboratory infrastructure required to implement PZA LPA, including proper sample handling (please also see section IV.). Feasibility for this test also hinges on the availability of an automated interpretation system, as it is difficult to interpret.

Part 2: Findings applicable to all 3 index tests

Sample Collection, Quality and Transport in the Context of New NAATs

The type of sample used as well as its collection, transport, and handling were all found to shape how the participants experienced TB diagnostics (please also see findings 9 and 10 in report by Engel et al., 2020).

Type: Extra-pulmonary samples are precious as the process of obtaining such samples is often not easy for the patient nor the doctor, so once the sample is obtained, clinicians will opt for the test that will most likely yield a result (i.e. phenotypic DST), as opposed to running it on a test that may come back inconclusive and require a re-run (molecular DST):

> *"[…] in these cases the intent is to take out as much of sample as possible and put up in one go, you see there the priority is not a molecular test. There the priority is to put up a culture. Suppose the sample is less, you cannot take out more sample, you put up for culture. culture is considered relative gold standard. Suppose I put up a molecular test and it doesn't give me anything, my entire effort for and investigation test has gone to waste. You can't take that chance" (IM-1)*

While the type of specimen guides which diagnostic a clinician will opt for (i.e. phenotypic DST for hard-to-obtain specimen), sample type may also determine where in the healthcare network the patient will access the diagnostic. For example, a clinician in South Africa discussed how local clinics are not used to or comfortable with inducing sputum or performing gastric washings (due to infection control) and therefore the patient has to be referred to a higher-level facility like a district hospital (SM-1). Similarly, a clinician in Moldova discussed that peripheral facilities are often not equipped to perform certain procedures to obtain extrapulmonary samples, such as a bronchoscopy (MM-1). As TB diagnostics become better able to detect TB from non-sputum-based samples, consideration must be placed on where they are positioned in the lab network. If they are placed too peripherally, they risk being underutilized at that level as certain specimens may only be collected at more centralized facilities.

Collection*:* Multiple participants discussed sample collection in the context of the current global pandemic. Due to the high priority health systems are placing on COVID-19 testing as well as heightened emphasis on infection control, participants reported that sample collection for TB testing has drastically decreased (IR-1, SM-1). Not only does this have important implications on TB case detection rates, it also has been found to influence TB diagnostic turnaround time and TB treatment initiation. A clinician in South Africa (SM-1) gave the example of individuals presenting to a clinic and through symptom screening, a decision is made to test for both TB and COVID-19. However, if the policy is that the patient should isolate at home until the COVID-19 results are ready, yet the patient tests positive MTB on GeneXpert within the average two-day TAT, TB treatment initiation could be delayed as the patient continues to isolate. Similarly, a clinician in Moldova (MM-1) discussed that as he waited for culture results, he would often perform bronchoscopies in order to obtain better quality sample from the infected locale (based on X-ray) to run on Xpert if previous sputum sample yielded negative results on microscopy and Xpert MTB/RIF. He noted that although this was an unpleasant procedure for the patient, it often would give him a sample that would test better than induced sputum. Due to COVID-19 and resulting infection control measures, he has not been able to perform bronchoscopies, and has reverted to waiting for culture, causing delays in clinical management (MM-1).

Transport*:* If an algorithm requires multiple samples at different times, from different collection points, as is the case when dealing with DR-TB, then an efficient sample transportation system is necessary. Many TB diagnostic algorithms require at least two samples to be collected: the first for TB diagnosis and second for DST. The relationship between number of samples needed, diagnostic TAT, and time to treatment initiation must be considered, especially in settings where there is no established sample collection and/or transportation mechanism (i.e. peripheral and rural settings; IP-2). Although Moldova is an example of a good sample transport system (ML-1, ML-2, MP-2, MM-1), it should be noted that the country – like many others – relies

on donor funding for various TB program activities, including sample transport. This raises concern about the sustainability of TB program funding and resources, a finding echoed by other qualitative research (Colvin 2015, deCamargo 2015, Cresswell 2014). As more diagnostics are adopted by programmes, funding towards systems that support these diagnostics (e.g. sample transportation) needs to be sustainable in order to get the most out of them.

***Handling**:* Sample handling plays an important role in the perceived reliability of LPAs and whole genome sequencing (IL-1, IL-2, IL-4, ML-2). If mishandled during preparation, the sample risks being contaminated and yielding inconclusive results on these molecular diagnostics. Here, participants cited good personnel skill, standardized operating procedures, and significant lab infrastructure as essential in reducing sample contamination in their laboratory (IL-2, IL-4).

<u>New diagnostics increase complexity when making clinical sense of results</u>

Laboratory staff often communicate diagnostic results to clinical staff in an automated manner via digital systems (e.g. TrakCare in South Africa). In some instances, laboratory staff and clinicians need to engage in further dialogue based on the results. For example, laboratory staff (or their pathologists) and clinicians may discuss changes in a patient's drug susceptibility, discordant results between diagnostic methods, or how to interpret results in relation to what concentration of a drug should be prescribed (SL-1, ML-1, ML-2). It is additional work for the clinician to constantly check the online system and see if culture results are available. This is illustrated by a clinician in South Africa who said;

> *"And also to try and, so one is the turnaround time does take a lot, and then sometimes things are messed, just because you know the clinics are really busy and if there is not a dedicated doctor to go through all of the results, sometimes the results are only picked up a lot later, and that's because you have to keep track of the culture. You know you have to keep on checking the culture, you know, not at four weeks it's still not out, not at six weeks it's still not out, eight weeks, then it's out, you know, and then it shows that there is a problem." (SM-1)*

Laboratory staff use information on which mutations confer high versus low-level drug resistance to discuss with clinicians about which drugs should be used or when they should be excluded (ML-1). Together, the clinician and laboratory staff may also talk about how certain doses have been working for the patient (ML-2). These discussions serve as an opportunity to explain why a strain may be susceptible in one test but not another or which drug the clinician should put confidence in based on the test results (MM-1).

New diagnostics will add to the complexity of discussing and interpreting results – work that relies on good two-way communication and established relationships between laboratory staff and clinicians. For example, new diagnostics could introduce additional complexity into discussions around drug concentration, dependent on their ability to detect high or low-level concentration resistance compared to culture DST. Here, we list two potential challenges in diagnostic communication, which need to be addressed while introducing new diagnostic methods:

1) Discrepant results may come in at different times (e.g. liquid culture vs. solid culture) (MM-1). Therefore, laboratory staff cannot indicate which result the clinician should go by, and in the case of discrepant results, this requires further communication between the laboratory staff and clinician on what diagnosis and treatment decisions should be made. This communication is sometimes delayed due to high laboratory workload (SM-1).

2) In some cases, doctors may be hesitant to take decisions on initiation of treatment for DR TB patients. For example, in Moldova, one participant stated that doctors at the district level are already hesitant to make decisions or changes to patient treatment regimens (which could technically be made on their own) without first consulting their national committee on DR TB patients (MM-1).

Therefore, implementation of new diagnostics must be accompanied with training for clinicians, to help them interpret results from new molecular tests and understand how this relates to treatment of a patient (ML-1) (see report Engel, finding 15). Furthermore, introduction of new diagnostics must be accompanied by guidelines and algorithms, which support clinicians and laboratories in communicating with each other in a timely manner,

such that they can discuss discordant results, and interpret laboratory results in the context of drug availability and patient history. Clinicians and laboratory staff also use patient progress to contextualize diagnostic results when there is uncertainty and use this to make treatment decisions (ML-2), but sometimes a patient is switched to another regimen based on test results despite clinical progress (SM-1). We suggest that policy-makers further incorporate clinical knowledge around patient progress to understand the reliability and clinical relevance of results from new diagnostics in practice.

Diagnostics and Health Equity

Social context presents barriers

It is well known that the difficult social contexts that many TB patients come from present various barriers to quick diagnosis and treatment initiation. It is therefore necessary to identify these barriers in order to promote the equity of and access to TB diagnostics and treatment. WHO notes in the 2020 Global TB report, "TB is a disease of poverty, and economic distress, vulnerability, marginalization, stigma and discrimination are often faced by people affected by TB". When discussing challenges in TB and MDR/XDR TB diagnosis, a research officer from a lab in India (IL-2) noted that the cost of testing and travelling places a burden on patients, especially if they are not being supported by the system. A programme officer from India (IP1) reiterated this, adding that even if they are being supported by the system, the far distances to travel becomes a barrier to DR TB treatment positioned at higher levels of the healthcare network. Similarly, an XDR TB survivor in India (IR1) discussed how he had to take a loan to afford his XDR TB treatment, while an XDR TB survivor from Moldova noted:

> *"The other challenges are connected to the economical material things, because you are losing job, and you have no actual resources to live, and treatment in such situation when you are not knowing, you have no stability for the next day, treatment goes somewhere…it is not a priority for people, that's true. Because when we are speaking about TB treatment its already acknowledged that you are paying only to the medical aspect…diagnosis and treatment is not enough for a successful treatment"*

A clinician in South Africa further gave an example of how social issues can intersect with the long TAT of TB diagnostics. Even though a particular test is considered a rapid diagnostic, such as Xpert MTB/RIF, the two-day TAT means that a patient will go back into the community without a diagnosis, and perhaps because of various social issues they face such as unemployment, substance abuse disorder, and food insecurity, the system may not be able to trace them once their result is ready. Closing this gap between presenting to the clinic and getting a diagnosis is a vital component of achieving the various global TB targets set in the SDGs, the End TB Strategy and the political declaration of the UN high-level meeting on TB.

In addition to access to fast, universal DST diagnosis and effective MDR/XDR TB treatment, access to information is a vital component of promoting equity in TB services. An XDR TB survivor (IR-1) noted that perhaps the biggest barrier that patients face in the Indian healthcare system is access to clear, comprehensible, and dependable information on what TB diagnostics are available to them and how to interpret results.

Need for a balance between diagnosis and treatment

There is reciprocal relationship between access to new diagnostics and access to new treatment. An XDR TB survivor from Moldova (MR-1) discussed that "*there should be always a balance between having good diagnostic tools and in the same time having options of treatment of drugs available because there is no use of a good diagnosis unless you have the necessary drugs to address the resistance*" while a clinician from India (IM-1) noted that "*Let us talk about the molecular test for linezolid, let us talk about any molecular test for bedaquilline and clofazimine. Think of it. These are newer drugs, they cannot be misused but obviously they have a cross resistance*". For a patient, it may be of little value to be able to test for resistance to TB medication, if that medication is not accessible to them, and for a clinician being able to prescribe a new TB treatment is made more valuable when he is able to test for its susceptibility early. Therefore, though it is important to improve access to treatment based on new diagnostics, it is equally important to improve access to diagnostics based on new treatment.

WHO guidelines and the speed of adoption

Finally, the speed at which WHO guidelines are changing does not match the speed at which many country programmes are able to implement the guidelines. This translates into differential access to new TB diagnostics and treatment at an inter-country level (i.e. between countries that can and cannot quickly keep up with the rapidly changing TB diagnostic environment MR-1) as well at an intra-country level (i.e. between patients who can and cannot afford the private health system that is better equipped to quickly adopt new diagnostics and policies; IL-1, IR-1).

Conclusions

The potential of new molecular diagnostics, which can quickly provide DST for a wider range of first and second line drugs, is immense. This could allow for quicker and more accurate treatment initiation. However, diagnostics do not have inherent value – they are only as good as their fit within the system in which they are working. Guidelines must take into consideration several challenges in order to support effective implementation of each abovementioned index test. Below, we list the challenges for each index test, as well as challenges related to samples, communicating diagnosis, and health equity.

1. Low-complexity automated NAATs (Xpert MTB/XDR Cartridge)
    - Cost specifically related to replacing/adjusting existing GeneXpert network
    - Deciding at what level of lab network it should be positioned

2. Medium-complexity automated NAATs
    - High cost of acquiring the platforms
    - Design of the system, some platforms might not physically fit in certain labs
    - Cost-efficiency related to number of samples versus controls per each test run; making sure the lab has enough samples to make each run cost-efficient

3. High-complexity hybridization-based NAATs (PZA LPA)
    - PZA DST is not routinely used in clinical settings, lack of clarity on PZA clinical relevance
    - LPA requires technical expertise and good laboratory infrastructure to reduce the risk of sample contamination
    - There is hesitancy among lab personnel to run LPAs on smear negative samples
    - Since PZA resistance is linked to multiple mutations, interpretation of PZA LPA bands is complicated and requires expertise and time

4. Sample
    - Potential lack of clinician confidence in using molecular diagnostics on hard-to-obtain samples
    - Collecting and transporting a good enough quality sample that can reliably run on a molecular test
    - Reliance on external or private funding for sample transport mechanisms and ensuring consistency in frequency and quality of sample transport system
    - Issues with sample contamination due to lack of technical skill and laboratory infrastructure

5. Communication
    - Clinicians facing difficulties getting a hold of laboratory personnel to discuss unclear results

- Hesitancy of clinicians to make decisions on MDR and XDR TB on their own
- Capacity of decentralized facilities to interpret and disseminate increased volumes of complex information produced by the tests

6. Equity
    - Barriers at program level: cost to acquire new expensive diagnostics as well as difficulties keeping up with WHO guidelines and subsequent training for clinicians and laboratory personnel
    - Barriers at patient level: lack of access to information and lack of access to treatments for which new diagnostics are testing
    - Lack of diagnostics available for newer treatments

These findings should be contextualized with COVID-19 in mind, which includes a move towards online training for clinicians, (lack of) access to testing and sample collection for patients, and increasing need for robust, flexible diagnostic systems in the face of new health challenges such as COVID-19 or increasing anti-microbial resistance.

Recommendations

1. **Low-complexity automated NAATs**: Although the Xpert MTB/XDR modular system allows for the decentralization of DST, placement of the XDR cartridge within healthcare/laboratory systems must consider:
    - The cost of 10-colour test
    - Sample load and treatment availability at a particular level/laboratory
    - Potential focus on MTB detection over DR TB detection at the primary care level
    - Expertise required to interpret the test

2. **Medium-complexity automated NAATs**: Placement within laboratory network should consider:
    - Minimal sample load to efficiently run tests
    - Physical specifications of systems as they relate to space available in the laboratory

3. **High-complexity hybridization-based NAATs (PZA LPA):** Guidelines for PZA LPA need to clearly address:
    - The clinical relevance of PZA resistance testing
    - When to order the PZA LPA in the testing algorithm
    - How to interpret and apply the results in the context of treatment, especially regarding which mutations confer or do not confer resistance (including continuing updates on which mutations do and do not confer resistance, especially as more surveillance data becomes available following implementation of the test)
    - Whether it can be run on smear-negative sputum samples
    - Provision of an auto-reader in order to reduce workload and address band-interpretation challenges

It is possible that increased testing for PZA resistance will improve understanding regarding the overall burden of PZA resistance. This could lead to a clearer justification as to why PZA DST is important. Furthermore, if PZA susceptibility is used routinely in treatment regimen decisions, this could translate into patient outcomes that either further justify (or challenge) the routine use of

PZA LPA (and PZA DST in general). This information should be monitored and taken into consideration for future guideline development.

All three index tests:

4. **Sample:**
    - Algorithms may need to consider how many samples are required from the patient, where they will be collected, and at which point in the patient journey the sample will be collected. This process should be streamlined where possible. Not doing so, there is risk of losing the patient before a diagnosis is made, or not being able to get additional samples from the patient as their conditions improves.
    - Guidelines should consider the volume of sample collected. For example if a patient tests positive for MTB and RIF resistant and that sample needs to be reflexed to Xpert MTB/XDR. Will the sample volume suffice to be run on Xpert MTB/XDR, LPA 1st and 2nd line, and be put up for liquid and solid culture, or will another specimen need to be collected?
    - Additionally, in order to avoid potential diagnostic delays, mechanisms that support TB diagnostics (i.e. sample transport and lab information systems) need to be well-funded and efficiently linked throughout the diagnostic network

5. **Communication:**
    - There should be guidance on how to develop algorithms which deal with discordant results from multiple kinds of tests within a shorter time span, as well as guidance on how to make clinical decisions regarding discordance (taking into consideration both laboratory and clinical knowledge)
    - Clear communication with clinicians on new diagnostic tests and training on how results should be applied within a clinical context.

6. **Equity:** Transparent and inclusive decision-making processes which include a variety of TB stakeholders in order to prioritize implementation of diagnostics tools that would be most useful in practice (taking into consideration the availability of and access to certain anti-TB medications)

7. **Future research:**

- Future research could explore digital laboratory information systems, how results are reported, and how this is affected by changing/increasing TB diagnostic information
- Future research should consider how diagnostic tests fit into existing decision-making structures, how programmes decentralize decision-making processes for treatment, especially as it relates to DR-TB, and whether peripheral healthcare providers are empowered to make treatment decisions that are currently made centrally

Proposed evidence for 'Evidence to Decision' tables in GRADE

Equity:

For patients, access to clear, comprehensible, and dependable information on what TB diagnostics are available to them and how to interpret results is a vital component to equity and represents an important barrier for patients in accessing care.

New treatment options need to be matched with new diagnostics: it is important to improve access to treatment based on new diagnostics, it is equally important to improve access to diagnostics for new treatment options.
The speed at which WHO guidelines are changing does not match the speed at which many country programmes are able to implement the guidelines. This translates into differential access to new TB diagnostics and treatment at an inter-country level (i.e. between countries that can and cannot quickly keep up with the rapidly changing TB diagnostic environment) as well at an intra-country level (i.e. between patients who can and cannot afford the private health system that is better equipped to quickly adopt new diagnostics and policies).

Acceptability:
Specific (infrastructure requirements, sample quality and volumes, communication between laboratory and clinicians) and general feasibility challenges (as identified in interview study and QES respectively), and accumulated delays risk undoing the added value/benefits as identified by the users (avoiding delays, drug resistant information). (combination QES Engel et al 2020 and interview study)
The **Xpert MTB/XDR cartridge (Low-complexity automated NAAT)** appears widely acceptable by laboratory staff and clinicians based on its simple user steps, familiarity of the system, and due to the amount of important information it provides. It addresses several preferences/values of laboratory staff and clinicians. It requires minimal user steps and the GeneXpert platform is a familiar system. The cartridge has a quicker turnaround time for first and second line drug susceptibility testing, compared to other available diagnostic methods. Respondents value faster TAT, the potential ability to reflex samples from the Xpert MTB/RIF to the Xpert MTB/XDR cartridge, and receiving information on multiple drugs as well as high or low level resistance simultaneously, as it could enable quicker diagnosis and optimized treatment for patients.
The automation of **medium-complexity NAATs**, which recognizes the high workload of laboratory staff, lends to the acceptability of these technologies. The physical size of the platform and how it fits into the laboratory space/workflow affect this acceptability (smaller footprint may be more acceptable). The number of samples run on the system is acceptable, if the platform is placed within a laboratory that receives a sufficient sample load to run the system. Medium-complexity automated NAATs address several preferences/values of clinicians and laboratory staff; it is faster than culture DST (like LPA or cartridge-based tests); has the advantage of being automated (unlike LPA); and gives additional clinically-relevant DR information e.g. high vs. low resistance (unlike the current GeneXpert MTB/RIF cartridge).
Hybridization-based Technology (PZA LPA): Acceptability of this method is dependent on how well it performs on different samples, as laboratory staff question how well LPA methods work on smear-negative samples. If samples first need to be cultured in order to run PZA LPA this may undermine the benefits of this method's quicker TAT compared to phenotypic DST for PZA. Acceptability also depends on how well it actually detects mutations specific to PZA resistance and clinicians and laboratory staff may require further clarification/justification in some settings as to why this specific DST drug test is being prioritized, as it is not currently part of routine DST. The PZA LPA addresses some preferences/values of laboratory staff and clinicians. It provides quicker results regarding PZA resistance, compared to other available methods (e.g. culture DST), can provide information on different concentration levels, and targets a drug that is widely used in first-line TB treatment.

Feasibility:
An **efficient sample transportation system, with sustainable funding mechanisms** is crucial for feasibility, especially if an algorithm requires multiple samples at different times, from different collection points, as is the case when dealing with DR-TB. If mishandled during preparation, the sample risks being contaminated and yielding inconclusive results on molecular diagnostics. Here, participants cited good personnel skill, standardized operating procedures, and significant laboratory infrastructure as essential in reducing sample contamination in their laboratory.

Introduction of new diagnostics must be accompanied by **guidelines and algorithms, which support clinicians and laboratories in communicating with each other**, such that they can discuss discordant results,

and interpret laboratory results in the context of drug availability, patient history, and patient progress on a current drug regimen.

Feasibility for the Xpert MTB/XDR Cartridge (Low-complexity automated NAAT) is challenged by the cost of the 10-colour system for this cartridge, and the value of diagnosing MTB over DR TB at primary care, makes it less feasible as a baseline test, though it would fit at a district or intermediate level laboratory. **The Xpert MTB/XDR Cartridge requires less user training** compared to other DST methods (such as LPA and culture), **making it more feasible to implement** compared to methods with more user steps and those methods which require significant additional training.
The feasibility of medium-complexity automated NAATs **is challenged by how/if the platform fits into the physical space of the laboratory** (considering bench size and weight of the platform). **A poorly functioning sample network challenges feasibility of implementing** medium-complexity automated NAATs and laboratory technicians voiced concerns over the quality of samples. Additional feasibility considerations for this method include ensuring clinicians and laboratory staff have time to communicate effectively regarding diagnostic results if the platform is centralized, while also ensuring the laboratory where it is placed is central enough to receive adequate numbers of samples to make the machine worth running.

Feasibility of PZA LPA is challenged by the significant training and laboratory infrastructure required to implement the test, including proper sample handling and quality sample. **Feasibility for this test also hinges on the availability of an automated interpretation system**, as it is difficult to interpret.

Acknowledgements

We would like to thank all participants for the time and insights they provided.

References

Albert H, Nathavitharana RR, Isaacs C, Pai M, Denkinger CM, Boehme CC. Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? *Eur Respir J*. 2016; 48: 516–25. https://doi.org/10.1183/13993003.00543-2016 PMID: 27418550

Cazabon, D., Pande, T., Kik, S., Van Gemert, W., Sohn, H., Denkinger, C., Qin, Z. Z., Waning, B., & Pai, M. (2018). Market penetration of Xpert MTB/RIF in high tuberculosis burden countries: A trend analysis from 2014 - 2016. *Gates open research*, 2, 35. https://doi.org/10.12688/gatesopenres.12842.2

Cepheid. (2020). Xpert® MTB/XDR. Retrieved November 28, 2020, from https://www.cepheid.com/en/tests/Critical-Infectious-Diseases/Xpert-MTB-XDR

Engel, N., Ochodo, E., Wanjiku Karanja, P., Schmidt, B.M., Janssen, R., Steingart, K.R., Oliver, S. (2020). User perspectives on cartridge-based nucleic acid amplification tests of low-complexity for tuberculosis and tuberculosis drug resistance.

Mwaura, M., Kao, K., Wambugu, J., Trollip, A., Sikhondze, W., Omesa, E., Dlamini, S., Mzizi, N., Dlamini, M., Sibandze, B., Dlamini, B., Albert, H., Dondorp, W., & Engel, N. (2020). Situating trade-offs: Stakeholder perspectives on overtreatment versus missed diagnosis in transition to Xpert MTB/RIF Ultra in Kenya and Swaziland. *PloS one*, *15*(2), e0228669. https://doi.org/10.1371/journal.pone.0228669

Stop TB Partnership. (2020). High Burden Countries. Retrieved November 29, 2020, from http://www.stoptb.org/countries/tbdata.asp

This report has been prepared for the WHO Global TB Programme. Do not distribute further.

World Health Organization. (2020). *Global Tuberculosis Report 2020*. Geneva: World Health Organization.

Web Annex D.12. LF-LAM for detecting active tuberculosis in people living with HIV: an updated systematic review

AUTHORS AFFILIATIONS

Stephanie Bjerrum[1,2], Ian Schiller[3], Nandini Dendukuri[3], Michael Eisenhut[4], Mikashmi Kohli[5], Ruvandhi R Nathavitharana[6], Alice A Zwerling[7], Claudia M Denkinger[8,9], Karen R Steingart[10], Maunank Shah[11]

[1]Department of Clinical Research, Research Unit of Infectious Diseases, University of Southern Denmark, Odense, Denmark
[2]OPEN, Odense Patient data Explorative Network, Odense University Hospital, Odense, Denmark
[3]McGill Global Health Programs and McGill International Tuberculosis Centre and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[4]Paediatric Department, Luton and Dunstable University Hospital NHS Foundation Trust, Luton, United Kingdom
[5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[6]Division of Infectious Diseases, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, USA
[7]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada
[8]FIND, Geneva, Switzerland
[9]Center of Infectious Diseases, University of Heidelberg, Germany
[10]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom
[11]Division of Infectious Diseases, Center for Tuberculosis Research, Johns Hopkins University, Baltimore, Maryland, USA

Corresponding author
Stephanie Bjerrum, MD PhD
Department of Clinical Research, Research Unit of Infectious Diseases
University of Southern Denmark
Winsløwparken 19, 3.sal, DK-5000 C Odense C
Denmark
sbjerrum@health.sdu.dk

Executive summary

The lateral flow urine lipoarabinomannan assay Alere Determine™ TB LAM Ag 'AlereLAM' is a commercially available point-of-care test that detects lipoarabinomannan, a lipopolysaccharide present in mycobacterial cell walls, in people with active tuberculosis (TB), including both pulmonary and extrapulmonary forms of disease. This systematic review summarizes the current literature on the accuracy of the AlereLAM for diagnosis of TB in people living with HIV as part of a World Health Organization process to develop updated guidelines for use of AlereLAM. AlereLAM is being considered as a diagnostic test that may be used in combination with existing tests for the diagnosis of HIV-associated TB. We report data on children separately from adults.

We identified 15 unique published studies that assessed the accuracy of AlereLAM in adults and integrated nine new studies identified since the original WHO and Cochrane reviews in 2015 and 2016. We classified studies that evaluated AlereLAM in participants with signs and symptoms of TB as 'studies with symptomatic participants' and studies that included both individuals with symptoms of TB and individuals without symptoms of TB (i.e. enrolled irrespective of symptoms) as 'studies with unselected participants'. All studies were performed in TB/HIV high burden countries. For this review, we report positive AlereLAM results in accordance with the manufacturer's updated recommendations for test interpretation (graded 1 to 4 based on band intensity). We estimated sensitivity and specificity at the grade 1 cut-off for positivity on the updated reference scale card, corresponding to grade 2 on the prior reference scale card with band intensities graded on a scale of 1 to 5. We performed all analyses with respect to a microbiological reference standard.

AlereLAM for TB diagnosis in HIV-positive adults with signs and symptoms of TB

Of the 15 included studies, eight studies reported accuracy data on AlereLAM for TB diagnosis among adults that presented with signs and symptoms of TB. Six of the studies contributed data partially or exclusively for inpatient settings. As assessed by QUADAS-2, six studies (75%) had high risk of bias in the patient selection domain, and seven studies (88%) had high risk of bias in the reference standard domain. Regarding applicability, we scored low concern for most studies in all domains. AlereLAM sensitivity and specificity varied with setting and CD4 cell count.

For all settings, AlereLAM pooled sensitivity and specificity (95% credible interval (CrI)) were 42% (31% to 55%) and 91% (85% to 95%), respectively (eight studies, 3449 participants (37% with TB); moderate-certainty evidence for sensitivity and low-certainty evidence for specificity).

Results of these studies indicate that in theory, for a population of 1000 people where 300 have microbiologically-confirmed TB, 189 would be AlereLAM-positive: of these, 63 (33%) would not have TB (false-positives); and 811 would be AlereLAM-negative: of these, 174 (21%) would have TB (false-negatives).

Stratified by setting, pooled sensitivity was 52% (40% to 64%) among inpatients versus 29% (17% to 47%) among outpatients. Pooled specificity was lower among inpatients, 87% (78% to 93%), versus 96% (91% to 99%) among outpatients.

Stratified by CD4 cell count, pooled sensitivity increased and specificity decreased with lower CD4 cell count. For all settings, in participants with CD4 ≤ 200 cells per μL pooled sensitivity was 45% (31% to 61%) versus 16% (8% to 31%) in participants with CD4 > 200 cells per μL. Pooled specificity was 89% (77% to 94%) for participants with CD4 ≤ 200 cells per μL and 94% (81% to 97%) for those with CD4 > 200 cells per μL. In participants with a CD4 ≤ 100 cells per μL pooled sensitivity was 54% (38% to 69%) versus 17% (10% to 27%) in participants with CD4 > 100 cells per μL. Pooled specificity was 88% (77% to 94%) in participants with a CD4 ≤ 100 cells per μL and 95% (89% to 98%) in participants with CD4 > 100 cells per μL. Pooled sensitivity in participants with CD4 between 101-199 cells per μL was 24% (14% to 38%).

AlereLAM for TB diagnosis in HIV-positive adults irrespective of signs and symptoms of TB

Of the 15 included studies, seven studies reported accuracy data on AlereLAM for TB diagnosis among unselected adults who may or may not have presented with TB symptoms at enrolment (i.e. enrolled

irrespective of signs and symptoms of TB). Studies were predominantly conducted in outpatient settings among patients with higher CD4 cell counts and lower TB prevalence compared with studies evaluating the test for TB diagnosis among exclusively symptomatic patients. Studies ranged from including 19% of participants with symptoms to 91%. As assessed by QUADAS-2, four studies (57%) had high risk of bias in the patient selection domain, and five studies (71%) in the reference standard domain. Regarding applicability, we scored low concern for all studies in all domains.

For all settings, AlereLAM pooled sensitivity and specificity were 35% (22% to 50%) and 95% (89% to 96%), respectively (seven studies, 3365 participants (13% with TB); moderate-certainty evidence for sensitivity and low-certainty evidence for specificity).

Results of these studies indicate that in theory, for a population of 1000 people where 100 have microbiologically-confirmed TB, 80 would be AlereLAM-positive: of these, 45 (56%) would not have TB (false-positives); and 920 would be AlereLAM-negative: of these, 65 (7%) would have TB (false-negatives).

Stratified by setting, pooled sensitivity was 62% (41% to 83%) among inpatients versus 31% (18% to 47%) among outpatients. Pooled specificity was lower among inpatients, 84% (48% to 96%) versus 95% (87% to 99%) for outpatients.

For all settings, stratified by CD4 cell count, in unselected participants with CD4 ≤ 200 cells per μL, AlereLAM pooled sensitivity and specificity were 26% (9% to 56%) and 96% (87% to 98%) (two studies). Pooled sensitivity in participants with a CD4 ≤ 100 cells per μL was 47% (30% to 64%) versus 20% (10% to 35%) in participants with CD4 > 100 cells per μL. Specificity was 90% (77% to 96%) in participants with a CD4 ≤ 100 cells per μL and 98% (95% to 99%) in participants with CD4 > 100 cells per μL. For other CD4 strata we had limited data.

Impact of AlereLAM on mortality

We identified two multi-site randomized controlled trials that included data on the impact of AlereLAM on mortality and other patient outcomes. Both trials were conducted in sub-Saharan Africa, with each including a study site in South Africa. Both trials involved hospitalized, HIV-positive patients, used the results of AlereLAM to guide therapy, and assessed all-cause mortality at eight weeks. We note that data stratified by CD4 count were limited. In the meta-analysis, the pooled risk ratio for mortality was 0.85 (0.76 to 0.94), that is, study participants undergoing AlereLAM testing had a 15% lower risk of mortality than participants undergoing routine TB diagnostic testing without AlereLAM; the absolute effect was 35 fewer deaths per 1,000 (from 14 fewer to 55 fewer deaths) (high-certainty evidence).

Association of AlereLAM and mortality

We identified 12 studies that had data on the association between AlereLAM positivity and mortality as part of post-hoc analyses within diagnostic accuracy studies (in which AlereLAM was not used for clinical decision making). The timing of mortality analysis, setting, use of TB therapy, and outcome measures to compare AlereLAM positive and AlereLAM negative patients differed across studies. In a descriptive analysis, 11 out of 12 of these studies suggested that there was an association of AlereLAM test positivity and mortality. Of importance, we note that these studies did not use results of AlereLAM to guide therapy.

AlereLAM studies in children

We identified three published studies of AlereLAM in children as the result of a broader search for studies in adults and children using the same inclusion criteria. The three studies involved a total of 266 HIV-positive children. One study enrolled children aged 14 years and less; one study enrolled children aged 12 years and less; and one study enrolled children aged 15 years and less. For the three studies, median age ranged from 24 months to 6.8 years. Two studies included HIV-positive children presumed to have TB with symptoms and one included HIV-positive children irrespective of TB signs and symptoms. One study was conducted in an outpatient setting, one in an inpatient setting, and one in both an inpatient and an outpatient setting. All three

studies took place in high TB/HIV burden countries in Africa. The prevalence of microbiologically-confirmed TB ranged from 7% to 40% in the studies.

Given the differences in population and setting, we did not perform meta-analyses and provide sensitivity and specificity estimates for individual studies. In all settings, including all children, sensitivity and specificity (95% CI) were 42% (15% to 72%) and 94% (73% to 100%), (30 participants, outpatient); 56% (21% to 86%) and 95% (90% to 98%), (130 participants, inpatient); and 43% (23% to 66%) and 80% (69% to 88%), (106 participants, both inpatient and outpatient).

Two studies provided data stratified by age group. In adolescents, AlereLAM sensitivities were 100% (3% to 100%) (four participants, inpatient) and 60% (15% to 95%) (nine participants, outpatient); in both studies, specificity was 100%. In children ≤ 5 years, sensitivities were 50% (7% to 93%) (95 participants, inpatient), and 25% (1% to 81%) (13 participants, outpatient); corresponding specificities were 93% (86% to 98%) and 89% (52% to 100%).

Authors' conclusions

We found that AlereLAM has lower sensitivity to detect TB in adults living with HIV than the internationally suggested target of minimum 65% overall for non-sputum based TB tests (WHO TTP 2014). This finding was consistent whether the test is used for diagnosis of TB among symptomatic participants (sensitivity of 42%) or unselected participants (sensitivity of 35%). The estimated sensitivity suggests that if AlereLAM were to be used alone, more than half of all TB cases would be missed. Although the estimated sensitivity is lower than the WHO target for non-sputum based TB tests, two randomised controlled trials implementing AlereLAM have demonstrated a mortality reduction and impact on other patient health outcomes when used in hospitalized HIV-positive adults.

The proposed role for the AlereLAM test is to be used in combination with other existing TB tests to assist TB diagnosis and possibly improve important outcomes among HIV-positive patients with advanced disease. The test does not require sputum collection and is not site-specific. Other favorable test characteristics include low-cost, rapidity (less than one hour), ease of use (does not require extensive sample preparation), and the fact that the test does not require electricity or special instruments and equipment (WHO TTP 2014).

Findings suggest that sensitivity increases with lower CD4 cell count and among inpatients regardless of the approach to enrolment of study participants (symptomatic versus unselected), but with a decrease in specificity. Overall estimates of specificity were approaching the internationally suggested target of 98% for non-sputum-based TB tests (WHO TTP 2014). Whether lower specificity among inpatients and individuals with lower CD4 can be attributed to misclassification of true positives as false positives due to an imperfect reference standard, or is due to other biological or environmental factors is unclear.

An increased number of studies were included in this updated review compared to the original review on LAM from 2015. However, we found considerable heterogeneity across studies and there were limited data for some sub-group analyses with respect to setting and CD4 count. Most studies used a lower quality reference standard where only sputum was microbiologically tested, and this may have led to misclassification of true-positive results as false-positive results (i.e. reduced specificity estimates). Many studies excluded participants unable to produce sputum, the target population expected to benefit the most from urine-based testing as they cannot have other sputum-based diagnostic testing.

All studies except one were conducted in sub-Saharan Africa, and we wish to underscore a concern about the applicability of the results on the whole outside of sub-Saharan Africa. We further consider the impact of AlereLAM to be affected by a number of factors, including the health care infrastructure and access to other diagnostic tests, prevalence of MDR-TB (which AlereLAM misses), and rates of empiric TB treatment. The results should, therefore, be interpreted with caution.

Concerning the accuracy of AlereLAM for TB in children living with HIV, there were too few studies and participants to draw conclusions.

Background

Tuberculosis (TB) remains the leading cause of hospitalization and in-hospital deaths among people living with HIV despite the increased access to antiretroviral treatment (ART) (Ford 2016). A systematic review of the prevalence of TB identified at autopsy suggests that, in resource-limited settings, TB is responsible for around 40% of all HIV-related deaths and that TB often was disseminated and undiagnosed at the time of death (Gupta 2015). Globally in 2017, only 51% of the estimated 10.0 million TB cases were notified among people living with HIV (WHO Global Report 2018). However, most death from TB is preventable if TB is detected early and effectively treated.

To improve TB case detection, new diagnostic tools and strategies for systematic screening of people living with HIV is a key component of the World Health Organization's (WHO) "End TB strategy" (WHO End TB 2014). Non-sputum-based point-of-care TB diagnostic tests are highly desired to narrow the diagnostic gap and ensure timely treatment (WHO TTP 2014). Desired characteristics of such a test would include minimal or non-invasive sample collection, short time to result (under one hour), and ability to implement the test without need for special instruments, electricity, or specimen preparation (WHO TTP 2014). Detection of mycobacterial antigen in urine has attracted great attention over time. Urine-based antigen testing would allow for a TB diagnosis that is non-site specific. Urine is further easy to collect and store, and lacks the infection control risks associated with sputum collection. Multiple platforms have been developed to detect lipoarabinomannan (LAM), initially as enzyme-linked immunosorbent (ELISA) assays that were evaluated in several clinical settings (Minion 2011). Later, the lateral flow assay, Alere Determine™ TB LAM Ag assay 'AlereLAM', was developed as a simple point-of-care test for diagnosis of active TB in people living with HIV. AlereLAM is commercially available, does not require access to special laboratory equipment, and produces a result after 25 minutes (Alere 2017), meeting many of the desired target product profile requirements (WHO TTP 2014).

The AlereLAM is recommended in the WHO Policy Guidance, "The use of lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis and screening of active tuberculosis in people living with HIV", published in 2015 (WHO Lipoarabinomannan Policy Guidance 2015). The guidance was informed by a review of evidence that was subsequently published in the original Cochrane Review of the AlereLAM (Shah 2016). The guidelines recommend that AlereLAM "may be used to assist in the diagnosis of TB in HIV-positive adult inpatients with signs and symptoms of TB (pulmonary/and/or extrapulmonary) and a CD4 cell count less than or equal to 100 cells per µL, or in people living with HIV who are 'seriously ill' regardless of CD4 count or if the CD4 count is unknown (WHO Lipoarabinomannan Policy Guidance 2015)". The recommendations also apply to HIV-positive outpatients and children with signs and symptoms of TB (pulmonary and/or extrapulmonary) based on the generalization of data from adult inpatients while acknowledging the limitation of available data (WHO Lipoarabinomannan Policy Guidance 2015). The WHO recommends that AlereLAM should not be used for general TB screening "owing to suboptimal sensitivity" (WHO Lipoarabinomannan Policy Guidance 2015). The guidelines further suggest that AlereLAM should be used in combination with existing tests, and not as a replacement test (to existing tests).

Of note, in 2018, preliminary performance characteristics of a second commercially developed lateral flow assay to detect LAM for the diagnosis of TB was announced based on data from frozen biobank specimens (Fujifilm SILVAMP TB LAM, Japan; FujiLAM) (Broger 2018). The test is projected to become commercially available in 2020.

The current systematic review includes published studies evaluating the commercially available AlereLAM assay for diagnosis of active TB disease (pulmonary and extrapulmonary TB) in people living with HIV. Since 2015, additional evidence for the use of AlereLAM has emerged. This updated systematic review will inform the WHO Guideline Development Group if there is evidence to update or modify recommendations on the use of AlereLAM.

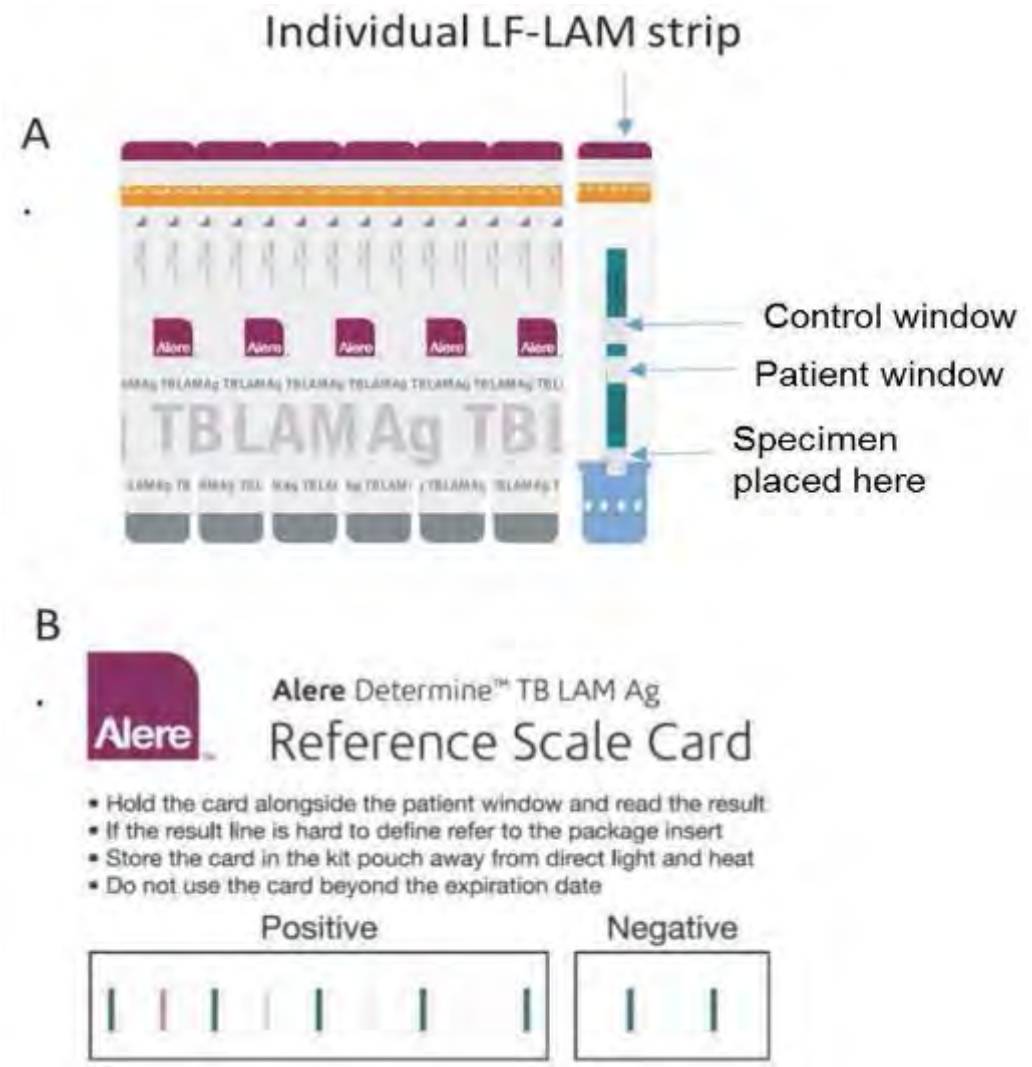
Index test

The urine-based lateral flow lipoarabinomannan assay AlereLAM is a commercially available point-of-care test for active TB (Alere Determine™ TB LAM Ag, Abbott, Palatine, IL, USA, previous Alere Inc., Waltham, MA, USA). AlereLAM is an immunocapture assay that detects LAM antigen in urine. LAM is a lipopolysaccharide present in mycobacterial cell walls (Brennan 2003), which is released from metabolically active or degenerating bacterial cells during TB disease (Briken 2004). LAM is detectable in urine of people with active TB disease and evaluated for both LAM ELISA and the lateral flow AlereLAM testing platforms (Peter 2010; Lawn 2012; Minion 2011; Shah 2016). The original Cochrane Review of AlereLAM (Shah 2016) and a meta-analysis of an earlier generation LAM ELISA test (Minion 2011) both demonstrated that the accuracy of urinary LAM detection was improved among people living with HIV with advanced immunosuppression. Several hypotheses may explain the higher sensitivity of urine LAM detection in people living with HIV including higher bacillary burden and antigen load (Shah 2010), greater likelihood of genitourinary tract TB involvement, and greater glomerular permeability to allow increased antigen levels in urine (Minion 2011; Lawn 2016). Based on current WHO guidelines, the role of the test can be characterized as a test to be used in combination with existing TB tests.

AlereLAM is performed manually by applying 60 µL of urine to the test strip and incubating at room temperature for 25 minutes (Alere 2017). See Figure 14. The strip is then inspected by eye. The intensity of any visible band on the test strip is graded by comparing it with the intensities of the bands on a manufacturer-supplied reference scale card. Of note, the reference scale was revised in January 2014. Prior to January 2014, the reference scale card included five bands (grade 1 representing a very low intensity band to grade 5 representing a high/dark intensity band). Some studies prior to January 2014 utilized grade 1 as the threshold for test positivity, while other studies utilized grade 2 as the positivity threshold. After January 2014, the manufacturer revised the reference scale card to have four reference bands, such that the band intensity for the new grade 1 corresponded to the band intensity for the previous grade 2. Under the current manufacturer recommendations (using the current 4 bands reference card), only bands that are grade 1 or higher are considered positive (Alere 2017). See Appendix 1. Reference card grading of Alere Determine™ TB LAM.

Figure 14. Alere Determine ™ TB LAM Ag tests, 'AlereLAM'

(A) Alere Determine™ TB LAM Ag tests. To the sample pad (white pad marked by the arrow symbols) 60 µL of urine is applied and visualized bands are read 25 minutes later. (B) Reference card accompanying test strips to 'grade' the test result and determine positivity.



Footnote: The reference scale card was changed in 2014. See Appendix 1. Reference card grading of Alere Determine™ TB LAM

PICO questions
We addressed the following PICO questions also listed in Appendix 2. PICO questions.

What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in all HIV-positive adults and children with signs and symptoms of TB?
in inpatient settings (adults, adolescents and older children)
in outpatient settings (adults, adolescents and older children)
all settings (adults, adolescents and older children)
in inpatient settings (children ≤ 5 years)
in outpatient settings (children ≤ 5 years)
all settings (children ≤ 5 years)

What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in all HIV-positive adults and children irrespective of signs and symptoms of TB?
in inpatient settings (adults, adolescents and older children)
in outpatient settings (adults, adolescents and older children)
all settings (adults, adolescents and older children)
in inpatient settings (children ≤ 5 years)
in outpatient settings (children ≤ 5 years)
all settings (children ≤ 5 years)

What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in adults with advanced HIV disease irrespective of signs and symptoms of TB?
in inpatient setting CD4 ≤ 200
in outpatient setting CD4 ≤ 200
in all settings CD4 ≤ 200
in inpatient setting CD4 ≤ 100
in outpatient setting CD4 ≤ 100
in all settings CD4 ≤ 100

Can the use of LF-LAM in HIV-positive adults reduce mortality associated with advanced HIV disease?
in all settings
in inpatient settings
in outpatient settings
in individuals with CD4 ≤ 200
in inpatient setting CD4 ≤ 200
in outpatient setting CD4 ≤ 200
in individuals with CD4 ≤ 100
in inpatient setting CD4 ≤ 100
in outpatient setting CD4 ≤ 100

Other questions: What is the cost and cost-effectiveness of LF-LAM implementation for TB diagnosis, based on review of the published literature?

People living with HIV are at increased risk of TB and may present with symptoms of TB but may also be asymptomatic or have symptoms not routinely associated with TB disease. To estimate accuracy in HIV-positive individuals with signs and symptoms of TB (PICO 1), we combined studies in which presentation with signs and symptoms suggestive of TB was an inclusion criterion and refer to these as ‘Studies with symptomatic participants’.

To estimate accuracy in HIV-positive adults irrespective of signs and symptoms of TB (PICO 2 and PICO 3), we combined studies that considered all HIV-positive individuals eligible to participate, including both

individuals with and individuals without symptoms of TB and refer to these as ‘Studies with unselected participants’.

We reviewed data related to patient-important outcomes, in particular, the impact of AlereLAM implementation on mortality (PICO 4).

A priori we wanted to investigate heterogeneity by clinical setting (inpatient versus outpatient) and by CD4 cell count (CD4 ≤ 200; CD4 ≤ 100) (PICO 1a-f; PICO 2a-f; PICO 3a-f; PICO 4a-i).

Throughout the report, we presented the diagnostic accuracy for AlereLAM in children and adolescents separately from adults.

Economic evaluations of AlereLAM for TB are reported in another document.

Methods

Criteria for considering studies for this review

Types of studies

We included primary studies evaluating the diagnostic accuracy of urine AlereLAM assay for the detection of active TB in people living with HIV and compared the index test results with a defined microbiological reference standard. We included studies from which we could extract true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) values.

Diagnostic studies for TB are largely cross-sectional in design but may include some clinical follow-up as part of patient classification. We included randomized controlled studies, cross-sectional studies and observational cohort studies and excluded case-control studies or other study designs. We excluded data reported only in abstracts, reviews, commentaries and editorial notes. We did not include unpublished data.

Participants

We included participants who were adults (15 years and older is considered 'adult' for purpose of TB surveillance) and HIV positive. We included studies in which there was a suspicion of TB among study participants based on the presence of signs and symptoms compatible with TB (studies with symptomatic participants), as well as studies that included participants who presented for medical care irrespective of signs and symptoms of TB (studies with unselected participants). Signs and symptoms of TB include cough, fever, weight loss, and night sweats. Participants who were known with active TB or taking anti-TB drug were not included.

Index tests

We included studies that evaluated Alere Determine™ TB LAM Ag test (Abbott, Palatine, IL, USA, previous Alere Inc., Waltham, MA, USA) ‘AlereLAM’ on urine samples. As of December 2018, AlereLAM was the only commercial lateral flow urine LAM assay available that had been evaluated in published studies. We included studies that evaluated the test at the manufacturer's recommended threshold for positivity i.e. grade 1 and above on the updated reference scale card with four band intensities graded on a scale of 1 to 4. For studies that used the prior reference scale card with band intensities graded on a scale of 1 to 5, we included those that evaluated the test at grade 2 and above corresponding to the current recommended positivity threshold. We excluded studies that did not use a positivity threshold corresponding to the manufacturer's recommendations. Results summarizing diagnostic accuracy at older thresholds (grade 1 on a scale of 1 to 5) can be found in the original review (Shah 2016).

Target conditions

The target condition was active TB disease among people living with HIV, which includes pulmonary and extrapulmonary TB.

Reference standards
We required studies to diagnose TB using the following microbiological reference standard.

'TB' is defined as a positive *M. tuberculosis* culture or Nucleic Acid Amplification Test (NAAT).
'Not TB ' is defined as a negative *M. tuberculosis* culture and NAAT (if performed).

NAAT tests included: Enhanced Amplified Mycobacterium Tuberculosis Direct Test (E-MTD, Gen-Probe, San Diego, USA); Amplicor *Mycobacterium tuberculosis* Test (Amplicor, Roche Diagnostics, Basel, Switzerland); COBAS® TaqMan® MTB Test (Roche Diagnostics); GenoType MTBDR*plus* (HAIN Lifesciences, Nehren, Germany); Xpert® MTB/RIF assay (Cepheid, Sunnyvale, USA); and Xpert® MTB/RIF Ultra.

For a microbiological reference standard, we considered a higher quality reference standard to be one in which two or more specimen types were evaluated for TB diagnosis in all participants as part of a standardised study algorithm. We considered a lower quality reference standard to be one in which only one specimen type was evaluated for TB diagnosis or if there was no algorithm defined to ensure a standardised approach for specimen collection and testing.

A microbiological reference standard, primarily culture, is considered the best reference standard. We expected all studies to obtain sputum specimens and some studies to obtain additional specimens for culture. However, the primary concern with relying on sputum culture alone is that TB diagnosis may be missed for the following reasons: people living with HIV may not be able to provide sputum specimens of sufficient quality; sputum bacillary load is typically low in people living with HIV; and a substantial proportion of people with HIV-associated TB cannot produce sputum at all (Lawn 2013a) or have extrapulmonary TB without pulmonary TB. This means that index test TPs may be misclassified as FPs by sputum culture. Therefore, when evaluating AlereLAM with respect to sputum culture, the number of FPs (classified as positive by the index test and negative by the reference test) may be increased and AlereLAM specificity may be underestimated (Lawn 2015). This misclassification may also lead to underestimation of sensitivity. Increasing the sensitivity of the reference standard by evaluating multiple specimens, including evaluating specimens from sites of disease for extrapulmonary TB, may reduce the number of cases of TB disease incorrectly classified as 'not TB' by culture or NAAT if performed.

In the original Cochrane Review, we additionally considered a ‘composite microbiological and clinical reference standard’ recognizing that microbiological reference standards alone may fail to detect TB in patients with TB disease. However, our original review found relatively little data using a composite reference standard; found heterogeneity in defining and applying composite reference standards; and found relatively modest impact on pooled estimates of sensitivity and specificity comparing microbiological and composite reference standards. Results assessing diagnostic accuracy against a composite reference standard can be found in the original review (Shah 2016).

Search methods for identification of studies
We performed literature searches up to 11 May 2018 in the following databases using the search terms reported in Appendix 3. Detailed search strategies: the Cochrane Infectious Diseases Group Specialized Register; MEDLINE (PubMed, from 1966); EMBASE (OVID, from 1947); Science Citation Index Expanded (SCI-EXPANDED, from 1900), Conference Proceedings Citation Index- Science (CPCI-S, from 1900), and BIOSIS Previews (from 1926), all three using the Web of Science platform; LILACS (BIREME, from 1982); and SCOPUS (from 1995). We also searched Clinicaltrials.gov and the search portal of the WHO International Clinical Trials Registry Platform (WHO ICTRP, www.who.int/trialsearch) to identify ongoing trials, and ProQuest Dissertations & Theses A&I (from 1861) to identify relevant dissertations. We included search results from the original review and re-evaluated previously included studies to determine if the studies met the refined inclusion criteria. We further examined reference lists of relevant reviews and studies and searched the WHO websites.

Selection of studies
We used Covidence systematic review software to manage the selection of studies (Covidence 2017). Two review authors (MS and SB) independently examined all titles and abstracts identified from the electronic search to determine potentially eligible studies. We obtained the full-text articles of these potentially eligible studies and the same two review authors independently assessed inclusion based on predefined inclusion and exclusion criteria. We resolved disagreements through discussion and, if necessary, consulted a third review author (KRS). We included studies from the original review if still eligible according to the predefined eligibility criteria.

Data extraction and management
We developed a standardized data extraction form and piloted the form on two of the included studies. Based on the pilot, we finalized the form. See Appendix 4. Data collection form. Then two review authors (MS and SB) independently extracted data from each included study on the following characteristics.

Author, publication year, study design, country(ies), clinical setting (outpatient or inpatient).
Participants: age, gender, HIV-status, CD4 count, TB history, clinical status (asymptomatic, symptomatic).
Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) items.
Cut-off used for determining a positive index test result and the reference card used.
Samples collected (sputum and/or extrapulmonary samples).
Reference standard(s) and the number of TB cases in the study.
Number of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) values for the index test.
Missing or unavailable test results.

We assigned country income status (high income, upper- and lower middle income, and low income) as classified by the World Bank (World Bank 2018/2019). In addition, we classified a country as being high burden or not high burden for TB/HIV according to the post-2015 era classification by the WHO (WHO Global Report 2018). We contacted study authors for clarifications on the AlereLAM positivity threshold used if data were missing.

We used REDCap electronic data capture tools (Harris 2009) hosted at OPEN, Odense Patient data Explorative Network, Odense University Hospital, Odense, Denmark (SDU Open) to collect and manage study data.

Assessment of methodological quality
We used the QUADAS-2 tool tailored to this review to assess the quality of the included studies (Whiting 2011; Appendix 6. QUADAS-2). QUADAS-2 consists of four domains: patient selection, index test, reference standard, and flow and timing (flow and timing domain includes differential verification of TB status for study participants). We assessed all domains for risk of bias and the first three domains for concerns regarding applicability. As recommended, we first developed the guidance on how to appraise the questions in each domain. Then, one review author (SB) piloted the tool with two of the included studies and finalized the QUADAS-2 tool. Two review authors (MS and SB) independently completed the QUADAS-2 judgements. We resolved disagreements through discussion or consulted a third review author (KRS).

Statistical analysis and data synthesis
We performed descriptive analyses of the characteristics of the included studies using Stata 15 (StataCorp 2017). We used the number of TPs, FPs, FNs, and TNs to calculate the individual study estimates of sensitivity and specificity and their 95% confidence intervals (CI). We presented individual study results graphically by plotting the estimates of sensitivity and specificity (and their 95% CIs) in forest plots using Review Manager (RevMan) (Review Manager).

We presented results at the current manufacturer reference scale card for test interpretation, with band intensities graded 1 to 4, and considered all test results at grade 1 and above as positive. The prior reference

scale card with five band intensities was used in the original Cochrane Review with grade 2 considered as positivity threshold that corresponds to the current grade 1 band intensity. See Appendix 1. Reference card grading of Alere Determine™ TB LAM. To allow consistent comparisons, we converted results from older studies that used the 'grade 2' threshold and treated these as 'grade 1' in the updated review. As such, analyses labelled at 'grade 2' in the original Cochrane Review are in this review considered according to the new manufacturer reference card as 'grade 1'. Studies in the original review that used the 'grade 1' threshold on the prior reference card were not included as this threshold is no longer recommended for determining test positivity.

We grouped the studies evaluating AlereLAM for: (I) diagnosis of TB in HIV-positive people with signs and symptoms of TB i.e. 'Studies with symptomatic participants' and (II) diagnosis of TB in HIV-positive people irrespective of signs and symptoms of TB i.e. 'Studies with unselected participants'.

When data were sufficient, we carried out meta-analyses to estimate AlereLAM pooled sensitivity and specificity with a bivariate random-effects model (Chu 2009; Reitsma 2005). This approach allowed us to calculate pooled sensitivity and specificity while dealing with potential sources of variation caused by: (1) imprecision of sensitivity and specificity estimates within individual studies; (2) correlation between sensitivity and specificity across studies; and (3) variation in sensitivity and specificity between studies.

We estimated all models using a Bayesian approach implemented using OpenBUGS (Lunn 2009). Under the Bayesian approach, all unknown parameters must be provided a prior distribution that defines the range of possible values of the parameter and the weight of each of those values, based on information external to the data. Because most meta-analyses involved few studies (eight or less), which could lead the model to be just identified, we chose to use low-information prior distributions for most parameters and a more informative prior on the between-study standard deviations which are particularly sensible in meta-analyses with few studies (Spiegelhalter 2004).

We defined prior distributions on the log-odds scale over the pooled sensitivity and specificity parameters, their corresponding between-study standard deviations (SDs) and the correlation between the sensitivities and specificities across studies. For the pooled log odds of the sensitivity or log odds of the specificity, we used a normal prior distribution with mean 0 and a variance of 4 (or a precision of 0.25). This corresponds to a roughly uniform distribution over the pooled sensitivity and pooled specificity on the probability scale. For the between-study precision we used a gamma distribution with a shape parameter of two and rate parameter of 0.5. This corresponds to a 95% prior credible interval (CrI) for the between-study SD in the log odds of sensitivity or log odds of specificity ranging from roughly 0.29 to 1.44, corresponding to moderate to high values of between-study heterogeneity. Covariance terms followed a uniform prior distribution whose upper and lower limits were determined by the sensitivity of the two tests. We have summarized the models we used (including the prior distributions) and the OpenBUGS programs we used to estimate them in Appendix 7. Statistical approach.

To study the sensitivity of our results to the choice of prior distributions given above, we considered alternative prior distributions that were less informative, which allowed a wider range of possible values. We increased the variance of the normal distributions over the pooled log odds of the sensitivity or specificity to 100. We used a uniform prior distribution ranging from zero to three over the between-study SD on the log odds scale. We found that the pooled estimates remained roughly the same with these alternative priors, though the posterior CrIs were wider, as expected. We combined information from the prior distribution with the likelihood of the observed data, in accordance with Bayes' theorem in the OpenBUGS program, which resulted in a sample from the posterior distribution of each unknown parameter. Using this sample, we calculated various descriptive statistics of interest. We estimated the median pooled sensitivity and specificity and their 95% CrI. The median or the 50% quantile is the value below which 50% of the posterior sample lies. We reported the median because the posterior distributions of some parameters may be skewed, and the median would be considered a better point estimate of the unknown parameter than the mean in such cases. The 95% CrI is the Bayesian equivalent of the classical (frequentist) 95% CI (we indicated 95% CI for

individual study estimates and 95% CrI for pooled study estimates as appropriate). The 95% CrI may be interpreted as an interval that has a 95% probability of capturing the true value of the unknown parameter given the observed data and the prior information.

In our original review we evaluated the incremental change in sensitivity and specificity when combining AlereLAM with smear microscopy or Xpert MTB/RIF (Shah 2016). We did not undertake analysis of incremental benefit in the current review as it was beyond the scope of this review, and data within published manuscripts was limited.

Approach to uninterpretable AlereLAM results
We excluded uninterpretable test results from the analyses for determination of sensitivity and specificity, but these were very few in numbers across studies.

Investigations of heterogeneity
Several PICO questions specifically sought to assess diagnostic accuracy among subgroups. A priori and when data were sufficient, we performed subgroup analyses with the following categorical covariates: clinical setting (inpatient versus outpatient) and CD4 count (CD4 ≤ 200, CD4 ≤ 100).

To further investigate heterogeneity, we performed additional subgroup analyses for CD4 strata CD4 101-200; CD4 > 200 and; CD4 > 100 as well as by TB prevalence. We investigated heterogeneity for the group of ‘studies with symptomatic participants’ separately from the group of ‘studies with unselected participants’.

Sensitivity analyses
We performed sensitivity analyses by limiting inclusion in the meta-analysis to the following.
Studies that avoided inappropriate exclusions, for example, studies that included participants who could not produce sputum. For this analysis we included studies that we scored as 'yes' for the QUADAS-2 question, "Did the study avoid inappropriate exclusions?" (low risk of bias for participant selection).
Studies with a higher quality reference standard, for example studies that included two or more specimen types. For this analysis, we included studies that we scored as ‘yes’ for the QUADAS-2 question, “Is the reference standard likely to correctly classify the target condition?” (low risk of bias for the reference standard).
Studies that used only fresh urine specimens for LAM testing
Studies initially categorized as ‘studies among unselected participants’ that included more than 80% of symptomatic participants were re-categorized as 'studies with symptomatic participants’. We conducted this analysis to explore the possibility that these studies represented a comparable population to the studies of symptomatic participants even though participants were not explicitly enrolled in the study on the basis of specific TB symptoms.

Additional analyses
Investigations of heterogeneity, all studies combined
We performed several additional post-hoc analyses to inform interpretation of findings. We assessed performance for the two groups of studies combined i.e. studies with symptomatic participants combined with studies with unselected participants. We did this analysis overall and stratified by inpatients, outpatients, and CD4 strata (CD4 ≤ 200; CD4 ≤ 100; CD4 > 200 and; CD4 > 100 cells per µL.)

Impact on mortality and other patient-important outcomes
Data that directly address the impact of test implementation on patient-important outcomes, such as mortality, are important for patients, decision makers, and the wider TB community. Our diagnostic test accuracy systematic review was not designed to answer questions on the impact of the test on patient outcomes, since a different methodology and separate search strategy would have been required. Nonetheless, we carried out additional efforts as follows. The primary reviewer authors (MS and SB) identified full text articles that included data on health impact. Another review author (RRN) examined all

15 articles included in this updated review, in addition to the other articles excluded during full text screening, and articles included in the original Cochrane Review (Shah 2016), to determine whether studies reported data on impact. We also looked for information on the association of test positivity and patient outcomes.

To evaluate impact data, we developed a standardized data extraction form and piloted this on two of the included studies. Based on the pilot, we finalized the form. See Appendix 5. Data collection form, impact data. Subsequently, one review author (RRN) extracted data from each included study on the following characteristics using Excel to collect and manage study data.

Author, publication year, study design, country(ies), clinical setting (outpatient or inpatient), number enrolled and analysed.
Participants: age, HIV-status, presence of symptoms (symptoms versus unselected).
Mode of mortality assessment, type of mortality (all-cause versus TB-related), timing of mortality assessment.
AlereLAM grade and use of old versus new reference card, timing of AlereLAM.
Mortality analysis metrics used (absolute risk reduction (ARR), (adjusted) Hazard Ratio (HR) or Kaplan Meier, (adjusted) odds ratio).
Comparator groups analysed.
Mortality in the intervention group, mortality in the control group (for randomized controlled trials).
Mortality in AlereLAM positive, mortality in AlereLAM negative.
Mortality in AlereLAM positive patients with confirmed TB, mortality in AlereLAM negative patients with confirmed TB.
Mortality in AlereLAM positive patients with inconclusive or non-TB diagnosis, mortality in AlereLAM negative patients with inconclusive or non-TB diagnosis.
Mortality data stratified by CD4 count.
Time to diagnosis, time to treatment.
Other outcomes assessed in the study.

For two randomized controlled trials, we assessed risk of bias using the Cochrane tool in RevMan (Review Manager). Then we narratively described the effect of AlereLAM implementation (that is, AlereLAM used versus AlereLAM not used to guide treatment) on patient-important outcomes including time to diagnosis and treatment, disease severity, and mortality. For mortality as a critical outcome, we performed a fixed-effect meta-analysis, including data from the two randomized trials, to estimate the pooled risk ratio (95% CI) for mortality. We thought it appropriate to use a fixed-effect approach because the estimates of the effect of the intervention in the different studies appeared similar, the differences between them being small enough to be explained by chance. However, a fixed-effect approach does not enable a measure of between-study heterogeneity.

Association between AlereLAM test positivity with patient-important outcomes
For diagnostic accuracy studies identified as having data on patient-important outcomes, we recorded whether there was an association between AlereLAM results and patient-important outcomes, including time to diagnosis and treatment, disease severity, and mortality. Of note, these studies did not use AlereLAM to guide treatment and are therefore not considered as impact studies.

Assessment of reporting bias
We did not carry out a formal assessment of publication bias using methods such as funnel plots or regression tests because such techniques have not been helpful for diagnostic test accuracy studies (Macaskill 2010).

Assessment of the certainty of the evidence

We assessed the certainty of evidence for intervention studies as recommended using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Balshem 2011). Although, the approach is similar for diagnostic studies, we describe it in detail here (Balshem 2011; Schünemann 2008; Schünemann 2016). As recommended, we rated the certainty of evidence as either high (not downgraded), moderate (downgraded by one level), low (downgraded by two levels), or very low (downgraded by more than two levels) based on five domains: risk of bias, indirectness, inconsistency, imprecision, and publication bias. For each outcome, the certainty of evidence started as high when there were high quality observational studies (cross-sectional or cohort studies) that enrolled participants with diagnostic uncertainty. If we found a reason for downgrading, we used our judgement to classify the reason as either serious (downgraded by one level) or very serious (downgraded by two levels).

Four review authors (SB, MS, ND, and KRS) discussed judgments and applied GRADE in the following way.

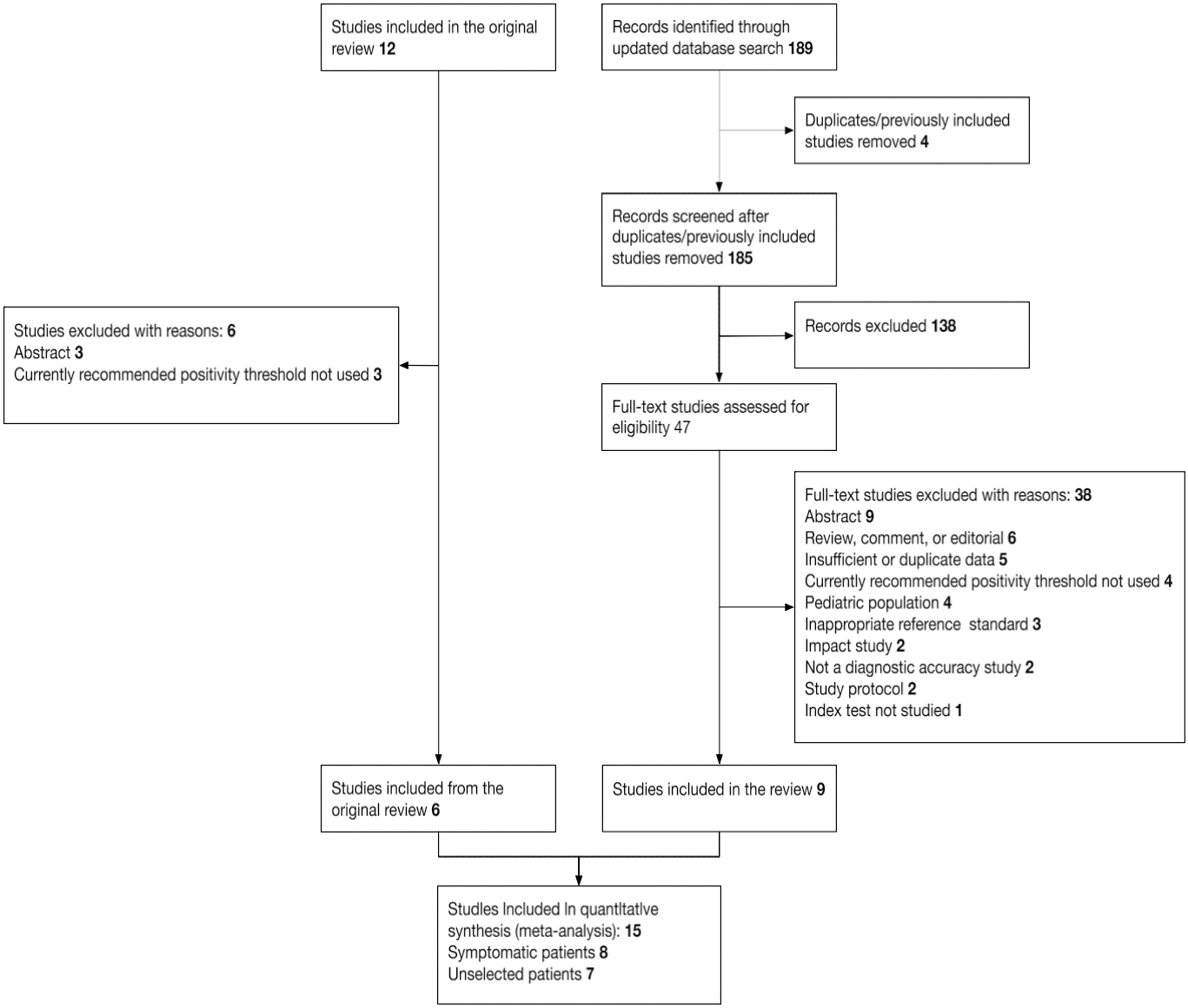
Risk of bias: we used QUADAS-2 to assess risk of bias.
Indirectness: We used QUADAS-2 for concerns of applicability and looked for important differences between the populations studied (for example, in the spectrum of disease), the setting, index test, and outcomes and asked are differences sufficient to lower certainty in results?
Inconsistency: GRADE recommends downgrading for unexplained inconsistency in sensitivity and specificity estimates. We carried out pre-specified analyses to investigate potential sources of heterogeneity and did not downgrade when we felt we could explain inconsistency in the accuracy estimates.
Imprecision: we considered a precise estimate to be one that would allow a clinically meaningful decision. We considered the width of the CrI, and asked ourselves, “Would we make a different decision if the lower or upper boundary of the CrI represented the truth?” In addition, we worked out projected ranges for TP, FN, TN, and FP for a given prevalence of TB and made judgements on imprecision from these calculations. We also considered whether the number of participants included in the analysis was less than the number generated by a conventional sample size calculation for a single adequately powered study (optimal information size).
Publication bias: we rated publication bias as undetected (not serious) for several reasons including the comprehensiveness of the literature search and extensive outreach to TB researchers to identify studies.

Results

Results of the search

We identified 15 unique studies that met the inclusion criteria of this review. We included data from six published manuscripts from the original WHO and Cochrane Review (WHO Lipoarabinomannan Policy Guidance 2015; Shah 2016) that met the refined inclusion criteria, and nine new studies identified in the updated search. Of six previously included studies, three were excluded because they did not use the currently recommended threshold for test positivity (Lawn 2012a; Balcha 2014; Drain 2014a); one abstract was included as an updated published manuscript (Lawn 2014a), one abstract remained unpublished (Andrews 2014), and one abstract was published but did not provide diagnostic accuracy data (Drain 2014c). Eight studies evaluated the accuracy of AlereLAM for TB diagnosis in participants with signs and symptoms suggestive of TB. Seven studies evaluated the accuracy of AlereLAM for diagnosis of unselected participants that may or may not have had TB signs and symptoms at enrolment. See Figure 2.

Figure 15. Flow of studies in the review.



Methodological quality of included studies

Figure 3 and Figure 4 show the quality assessment of the 15 included studies.

Studies with symptomatic participants

Eight studies were included that evaluated AlereLAM for TB diagnosis among symptomatic participants suspected of TB. In the patient selection domain, we considered six studies (75%) to be at high risk of bias because: (1) the study excluded all smear-positive participants (Drain 2016) (2) the studies excluded participants who could not expectorate or produce sputum despite sputum induction (Drain 2016; Nakiyingi 2014; Peter 2015); (3) the study excluded participants if they did not have a full set of complement reference standard results i.e. had any sample with a missing Xpert MTB/RIF result or a contaminated culture result in the absence of a positive result (Huerga 2017); (4) the study only included patients suspected of extrapulmonary TB and excluded patients suspected of pulmonary TB (Juma 2017); (5) the study only included participants with pericardial effusion and suspected TB and excluded participants suspected of other forms of TB (Pandie 2016). All studies were cross-sectional, cohort or randomized controlled studies. Regarding applicability, seven studies (88%) had low concern in the patient selection domain because the studies included the appropriate participants and settings. We judged one study (12%) to have high concern for applicability as the study participants did not resemble people with presumed HIV/TB co-infection i.e. participants were smear-negative HIV-positive and HIV-negative patients with a Karnofsky Performance score < 50 (Drain 2016).

In the index test domain, we judged one study (12%) at high risk of bias with a high concern of applicability as the study used grade 2 (on the current reference scale card) as the test positivity threshold, as opposed to the current manufacturer recommendation to use grade 1 (on the updated reference card) to define test positivity (Juma 2017). The remaining studies all used the recommended threshold for positivity and interpreted the test without knowledge of the results of the reference standard, and we considered them to have low concern for applicability.

In the reference standard domain, we considered seven studies (88%) to be at high risk of bias because: (1) the studies did not include testing of any extrapulmonary specimens (Drain 2016, Peter 2015); (2) the study did not include testing of any respiratory samples (Juma 2017); (3) the study only tested respiratory samples for some of the participants (Pandie 2016); (4) the study only tested extrapulmonary specimens in addition to respiratory samples for some of the participants (Huerga 2017); (5) health providers selected the sites for testing based on their own clinical suspicion (Peter 2012; Peter 2016). We deemed three studies at high concern for applicability as they lacked a study or protocol directed testing (Peter 2012; Peter 2016; Pandie 2016). In these studies, health providers selected the sites for testing based on their own clinical suspicion, and it was unclear if their choice of reference standard would correctly classify TB.

In the flow and timing domain, we considered four studies (50%) to be at high risk of bias because not all participants received the same reference standard (Huerga 2017; Peter 2012) or because not all participants were included in the two-by-two tables (Pandie 2016; Huerga 2017). We judged the remaining studies to be at low risk of bias because all participants received the index test and the same reference standard, and none of the participants enrolled in the studies were excluded from analysis.

Studies with unselected participants

Seven studies contributed data for the purpose of evaluating AlereLAM for TB diagnosis among unselected participants that may or may not have TB symptoms (Figure 4). In the patient selection domain, we considered four studies (57%) to be at high risk of bias because these studies excluded participants who could not expectorate or produce sputum samples (Bjerrum 2015; Floridia 2017 LaCourse 2016; Drain 2015). All studies were cross-sectional or cohort studies. Regarding applicability, we judged that all studies (100%) included the appropriate participants and settings.

In the index test domain, we considered all studies at low risk of bias as all studies used AlereLAM, pre-specified the grade used for positivity, and interpreted the test at the recommended positivity threshold without knowledge of the results of the reference standard. We considered the test conduct and interpretation in all studies to be applicable.

In the reference standard domain, we considered five studies (71%) to be at high risk of bias because these studies did not include microbiological testing on extrapulmonary specimens (Bjerrum 2015; Drain 2015; Floridia 2017; LaCourse 2016; Thit 2017). We judged these studies to be of low concern in terms of applicability. In one study it was unclear if the reference standard results was interpreted without knowledge of the index test result and we judge an unclear concern of applicability.

In the flow and timing domain, we considered two studies (29%) to be at high risk of bias because the study collected specimens for index and reference standard tests up to six months apart (Hanifa 2016; Thit 2017) and Hanifa 2016 excluded clinical TB cases from analysis. We considered the remaining five studies (71%) to be of low risk of bias because all participants received the index test and the same reference standard, and no participants enrolled were excluded from the two-by-two table.

Figure 16. Risk of bias and applicability concerns graph.
Review authors' judgements about each domain presented as percentages across included studies.

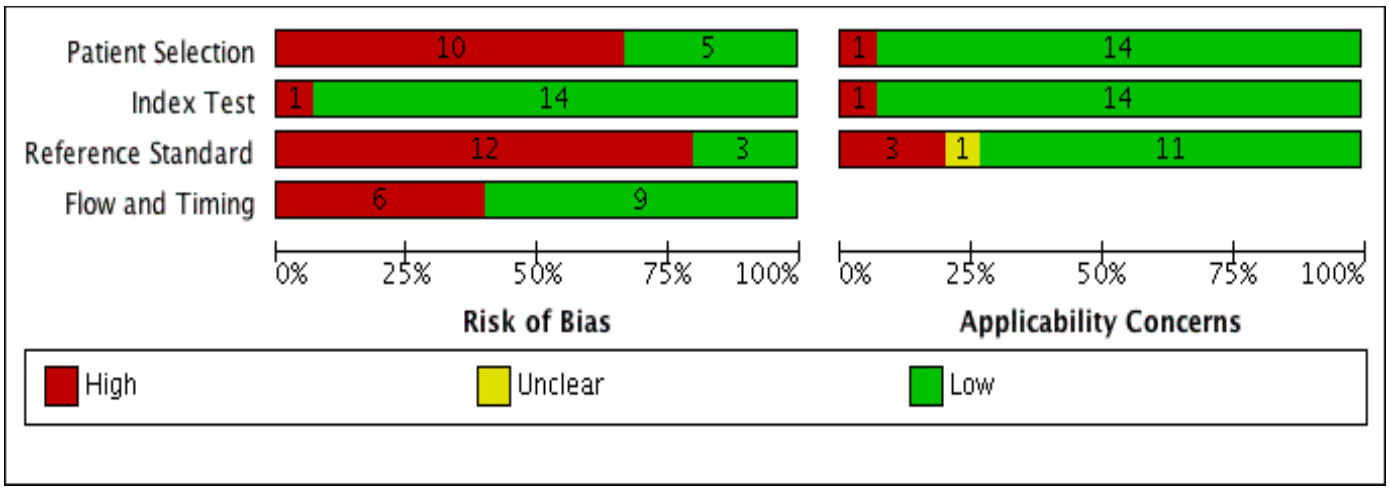
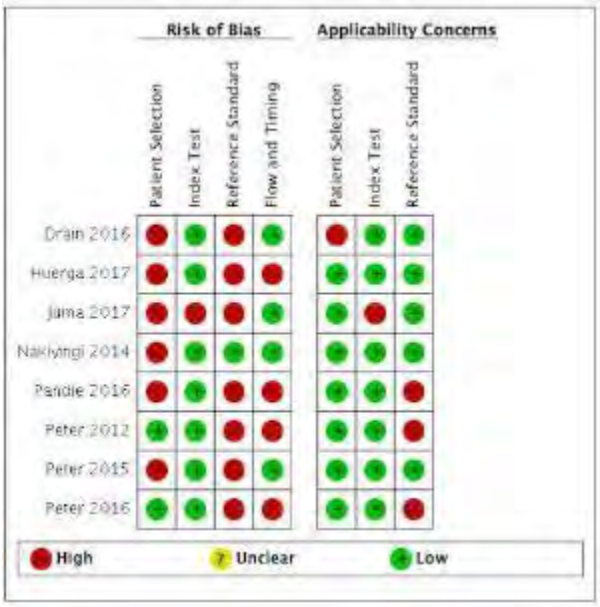


Figure 17. Risk of bias and applicability concerns summary.
Review authors' judgements about each domain for each included study. (A) Studies with symptomatic participants. (B) Studies with unselected participants.



Findings

The 15 included studies involved 6814 participants, 1761 (26%) with TB. Eight of the studies evaluated the accuracy of AlereLAM for TB diagnosis in participants with signs and symptoms suggestive of TB involving 3449 participants, 1277 (37%) with TB. Seven studies evaluated the accuracy of AlereLAM for diagnosis of unselected participants that may or may not have had TB signs and symptoms at enrolment involving 3365 participants, 439 (13%) with TB.

All studies were performed in high TB/HIV burden countries and classified as low-income or middle-income countries. We noted substantial differences in the studies for the following characteristics: type of study ('studies with symptomatic participants' and ‘studies with unselected participants'); setting (inpatients versus outpatients); median CD4 cell count; TB prevalence; inclusion and exclusion of participants based on whether or not they could produce sputa; and whether patients were evaluated for pulmonary TB, extrapulmonary TB or both. The key study characteristics are summarised by study in Appendix 8: Characteristics.

Most studies reported that a valid AlereLAM result was obtained on the first attempt for all tests. Few uninterpretable test results (< 1%) were reported in three studies (Peter 2012; Peter 2015; Peter 2016).

Table 23 presents pooled sensitivity and specificity results for AlereLAM against a microbiological reference standard grouped by the type of study 'TB diagnosis among symptomatic participants' and 'TB diagnosis among unselected participants’.

Table 23. AlereLAM pooled sensitivity and specificity for TB diagnosis, by study group

Type of analysis	Symptomatic participants				Unselected participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
Overall accuracy	8 studies (3449)	1277 (37%)	42% (31 to 55)	91% (85 to 95)	7 studies (3365)	432 (13%)	35% (22 to 50)	95% (89 to 98)
By setting								
Inpatient	6 studies (2253)	868 (39%)	52% (40 to 64)	87% (78 to 93)	3 studies (537)	159 (30%)	62% (41 to 83)	84% (48 to 96)
Outpatient	4 studies (1196)	409 (34%)	29% (17 to 47)	96% (91 to 99)	6 studies (2828)	273 (10%)	31% (18 to 47)	95% (87 to 99)
By CD4 cell								
CD4 > 200	3 studies (738)	163 (22%)	16% (8 to 31)	94% (81 to 97)	1 study[a] (156)	11 (7%)	Not applicable	Not applicable
CD4 ≤ 200	4 studies (1825)	722 (40%)	45% (31 to 61)	89% (77 to 94)	2 studies (706)	82 (12%)	26% (9 to 56)	96% (87 to 98)
CD4 > 100	4 studies (1519)	425 (28%)	17% (10 to 27)	95% (89 to 98)	4 studies (952)	115 (12%)	20% (10 to 35)	98% (95 to 99)
CD4 ≤ 100	4 studies (1239)	512 (41%)	54% (38 to 69)	88% (77 to 94)	3 studies (417)	130 (31%)	47% (40 to 64)	90% (77 to 96)
CD4 101-200	4 studies (586)	210 (36%)	24% (14 to 38)	90% (77 to 96)	1 study[b] (103)	13 (13%)	Not applicable	Not applicable
By CD4 and setting								
CD4 ≤ 200 inpatients	2 studies (1009)	348 (34%)	54% (34 to 73)	80% (58 to 91)	1 study[c] (54)	14 (26%)	Not applicable	Not applicable
CD4 ≤ 100 inpatients	2 studies (734)	270 (37%)	61% (40 to 78)	81% (61 to 91)	2 studies (200)	84 (42%)	57% (33 to 79)	90% (69 to 97)
CD4 101-200 inpatients	2 studies (275)	78 (28%)	32% (16 to 57)	81% (55 to 92)	1 study[d] (9)	4 (44%)	Not applicable	Not applicable
CD4 ≤ 200 outpatients	1 study[f] (249)	97 (39%)	Not applicable	Not applicable	2 studies (652)	68 (10%)	21% (8 to 48)	96% (89 to 99)
CD4 ≤ 100 outpatients	1 study[g] (121)	48 (40%)	Not applicable	Not applicable	2 studies (217)	46 (21%)	40% (20 to 64)	87% (68 to 94)
CD4 101-200 outpatients	1 study[h] (128)	51 (40%)	Not applicable	Not applicable	1 study[e] (94)	9 (10%)	Not applicable	Not applicable

Abbreviations: Crl: credible interval; AlereLAM: Alere Determine™ TB lipoarabinomannan assay; TB: tuberculosis.
[a] Bjerrum 2015, Sensitivity 27% (6% to 61%); Specificity 99% (96% to 100%); [b] Bjerrum 2015, Sensitivity 38% (14% to 68%); Specificity 99% (94% to 100%);[c] Bjerrum 2015, Sensitivity 64% (35% to 87%); Specificity 82% (67% to 93%); [d] Bjerrum 2015, Sensitivity 75% (19% to 99%); Specificity 100% (48% to 100%); [e] Bjerrum 2015, Sensitivity 22% (3% to 60%); Specificity 99% (94% to 100%); [f]Peter 2015, Sensitivity 24% (16% to 33%); Specificity 94% (89% to 97%); [g]Peter 2015, Sensitivity 30% (18% to 46%); Specificity 93% (85% to 98%); [h]Peter 2015, Sensitivity 18% (8% to 31%); Specificity 95% (87% to 99%).

PICO 1: What is the diagnostic accuracy of AlereLAM for TB diagnosis in HIV-positive adults with signs and symptoms of TB?

Of the 15 included studies, eight evaluated the accuracy of AlereLAM for TB diagnosis in participants with signs and symptoms suggestive of TB. The suggestive signs and symptoms of TB varied from study to study, but were often based on any of cough, fever, weight loss, or night sweats. The TB prevalence ranged from 29% to 63%. Two studies were conducted exclusively among patients with presumed extrapulmonary TB:

Juma 2017 and Pandie 2016. Four studies were conducted exclusively in an inpatient setting; two studies exclusively in an outpatient setting, and two studies in both inpatient and outpatient settings. The median CD4 cell count ranged from 81 to 210 cells per µL across the eight studies, lower in studies evaluating inpatients (median CD4 between 81-139 cells per µL) compared to studies evaluating outpatients (median CD4 was 168-210 cells per µL). See Appendix 8. Characteristics of Included Studies.

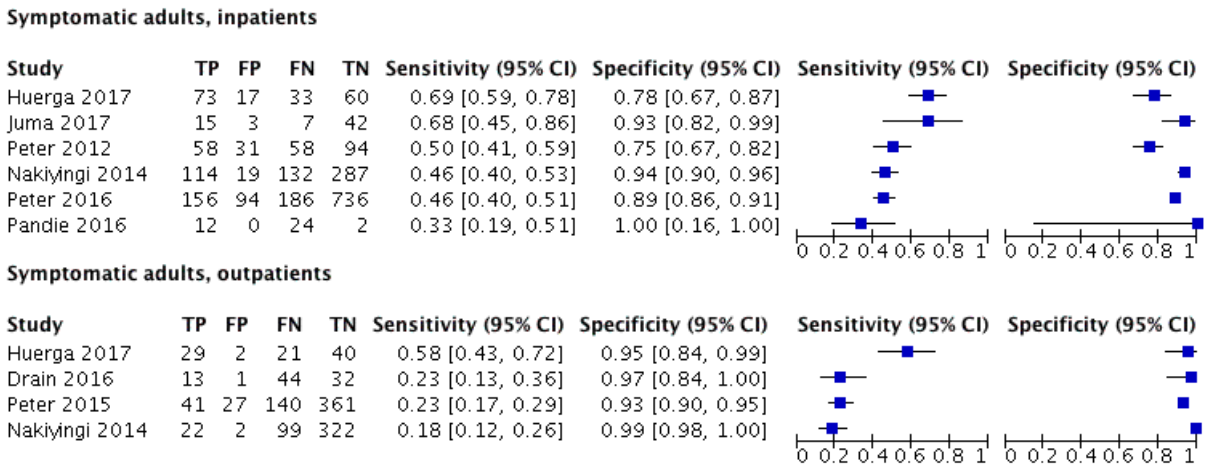
Results for children are presented in Appendix 9: Diagnostic accuracy of AlereLAM among HIV-positive children, summary.

PICO 1a: Accuracy, in inpatient settings
Six studies were conducted among inpatients involving 2253 participants, 868 (39%) with TB (Huerga 2017;Juma 2017; Nakiyingi 2014; Pandie 2016; Peter 2012; Peter 2016). Sensitivity estimates ranged from 33% to 69% and specificity estimates ranged from 75% to 100%. The pooled sensitivity and specificity (95% CrI) among inpatients were 52% (40% to 64%) and 87% (78% to 93%). See Figure 18. The highest sensitivity (69%) was reported by Huerga 2017 with a relatively low specificity (78%). This study included inpatients who were severely ill or with CD4 < 200 cell per µL (median CD4 109) and low BMI. The study did not include microbiological or histological evaluation of extrapulmonary specimens for TB in their reference standard, which may have led to LAM-positive participants with extrapulmonary TB being misclassified as 'false-positive' and lowered specificity. Pandie 2016 reported the lowest sensitivity (33%) and a specificity of 100%. This study differed from others by evaluating accuracy only for pericardial TB. The authors excluded a number of participants in the analysis for unknown reasons and reported TN as '2' and FP as '0' that may have inflated specificity. In the original review, the pooled sensitivity and specificity among inpatients were 53% (38% to 70%) and 90 % (73% to 96%) (four studies, 1299 participants), (Shah 2016).

PICO 1b: Accuracy, in outpatient settings
Four studies were conducted among outpatients involving 1196 participants, 409 (34%) with TB (Drain 2016; Huerga 2017; Nakiyingi 2014; Peter 2015). Sensitivity estimates ranged from 18% to 58% and specificity estimates ranged from 93% to 99%. The highest sensitivity (58%) was reported by Huerga 2017 where outpatients included were severely ill, or had a CD4 < 200 cell per µL, or a low Body Mass Index below 17Kg/m^2. Pooled sensitivity and specificity were 29% (17% to 47%) and 96% (91% to 99%). See Figure 18.

Pooled estimates were not previously calculated for outpatients due to lack of data in the original Cochrane Review (Shah 2016).

Figure 18. Diagnostic accuracy in adults with signs and symptoms, by health care setting
Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, by health care setting.



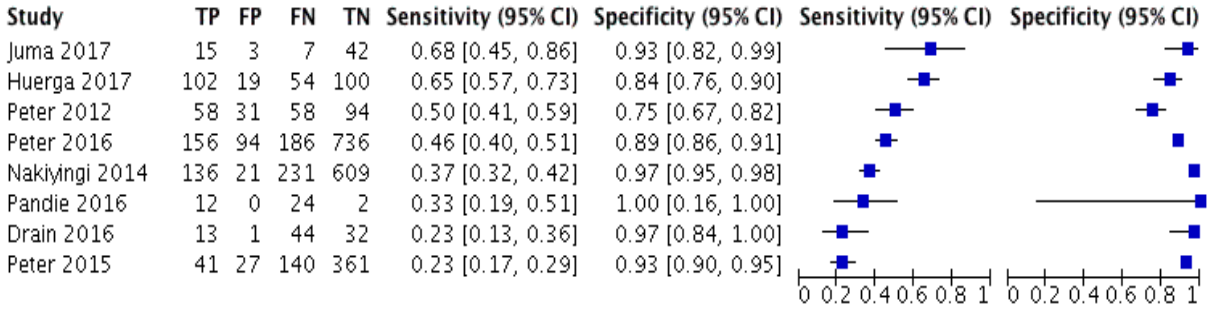
Type of analysis	Symptomatic participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
Inpatient	6 studies (2253)	868 (39%)	52% (40 to 64)	87% (78 to 93)
Outpatient	4 studies (1196)	409 (34%)	29% (17 to 47)	96% (91 to 99)

PICO 1c: Overall accuracy, all settings

For the analysis of the overall accuracy of AlereLAM in HIV-positive adults with signs and symptoms of TB, eight studies provided data for 3449 participants, including 1277 (37%) TB patients, (Drain 2016; Huerga 2017; Juma 2017; Pandie 2016; Peter 2016; Nakiyingi 2014; Peter 2012; Peter 2015). Sensitivity estimates ranged from 23% to 68%, and specificity estimates from 75% to 100%. The pooled sensitivity and specificity (95% CrI) were 42% (31% to 55%) and 91% (85% to 95%). See Figure 19. Juma 2017 evaluated diagnostic accuracy for extrapulmonary TB (all forms) exclusively and had sensitivity of 68%. Pandie 2016 evaluated accuracy for pericardial TB and found sensitivity of 33%. Sensitivity was lowest in the studies by Peter 2015 and Drain 2016. Differences between these studies and the other studies in this analysis were the setting (outpatient only), focus on pulmonary TB (no extrapulmonary samples were taken), and exclusion of participants unable to produce sputum. In particular, Drain 2016 included smear-negative participants with presumed TB and a small number of HIV-negative participants. In addition, this study excluded participants with a low Karnofsky score in order to target relatively well outpatients, where smear-negative TB disease is often seen. Specificity was lowest for Peter 2012, a study that included only inpatients and differed from other studies in that both sputum and non-sputum-based sampling was performed at the discretion of the attending clinical team and not study directed. Pandie 2016 reported a specificity of 100%, but excluded participants from specificity analysis as mentioned above. For comparison, the pooled sensitivity and specificity in the original review were 45% (29% to 63%) and 92% (80% to 97%) based on five studies and 2313 participants (Shah 2016).

Figure 19. Diagnostic accuracy in adults with signs and symptoms, all settings

Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, overall.



Type of analysis	Symptomatic participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
Overall accuracy	8 studies (3449)	1277 (37%)	42% (31 to 55)	91% (85 to 95)

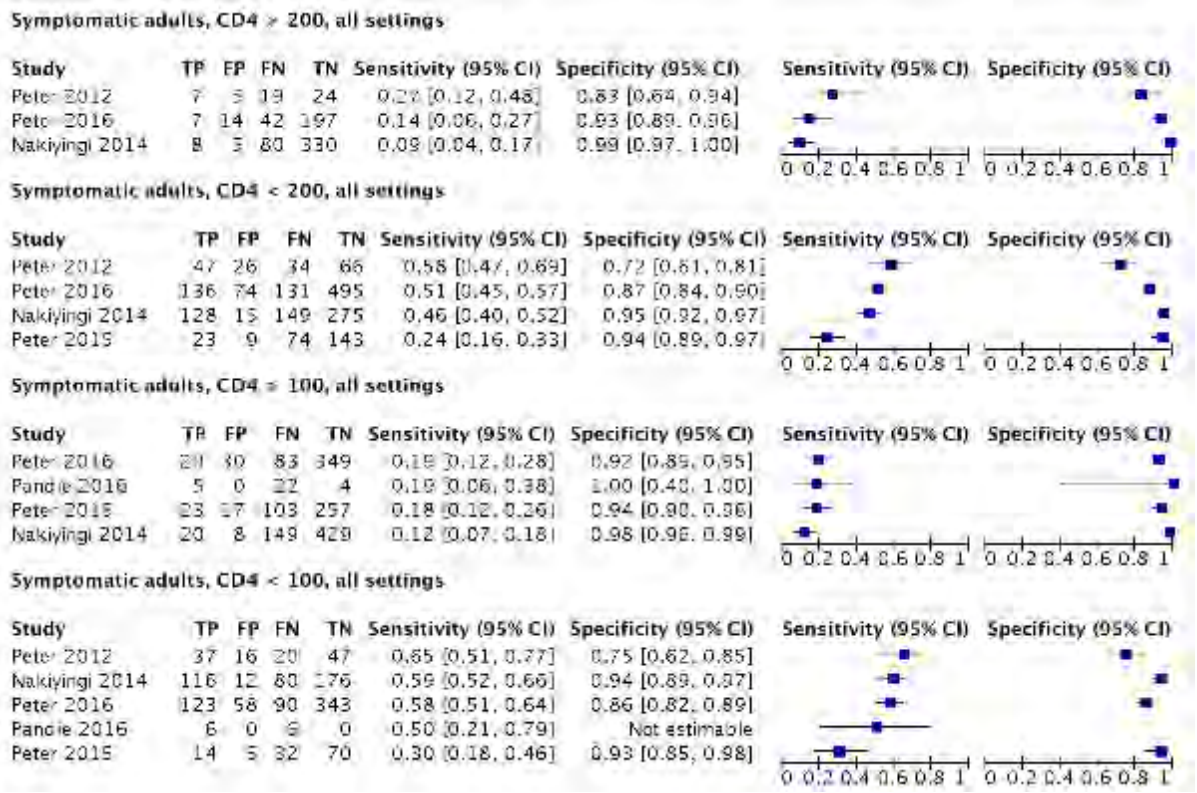
Additional investigations of heterogeneity

CD4 count

Accuracy stratified by CD4 > 200 cells per µL and ≤ 200 cells per µL

Three studies evaluated participants with CD4 > 200 cells per µL, (Nakiyingi 2014; Peter 2012; Peter 2016). Sensitivity estimates ranged from 9% to 27% and specificity estimates ranged from 83% to 99%. In the four studies that evaluated participants with CD4 ≤ 200 cells per µL, sensitivity estimates ranged from 24% to 58% and specificity estimates ranged from 72% to 95% (Nakiyingi 2014; Peter 2012; Peter 2015; Peter 2016). See Figure 20. The pooled sensitivity (95% CrI) was higher among participants with CD4 ≤ 200 cells per µL at 45% (31% to 61%) (1825 participants; 40% with TB) versus 16% (8% to 31%) among those with CD4 > 200 cells per µL (738 participants: 22% with TB). The pooled specificity was 89% (77% to 94%) for participants with CD4 ≤ 200 cells per µL and 94% (81% to 97%) for those with CD4 > 200 cells per µL. When we limited the analysis to studies involving inpatients with CD4 ≤ 200 cells per µL, the pooled sensitivity and specificity were 54% (34% to 73%) and 80% (58% to 91%) (two studies, 1009 participants; 34% with TB (Peter 2012; Peter 2016). Only one study reported data for outpatients with CD4 ≤ 200 cells per µL Peter 2015.

Figure 20. Diagnostic accuracy in adults with signs and symptoms, by CD4 count.
Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, by CD4.



Type of analysis	Symptomatic participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
CD4 > 200	3 studies (738)	163 (22%)	16% (8 to 31)	94% (81 to 97)
CD4 ≤ 200	4 studies (1825)	722 (40%)	45% (31 to 61)	89% (77 to 94)
CD4 > 100	4 studies (1519)	425 (28%)	17% (10 to 27)	95% (89 to 98)
CD4 ≤ 100	4 studies (1239)	512 (41%)	54% (38 to 69)	88% (77 to 94)

Accuracy stratified by CD4 > 100 cells per µL and ≤ 100 cells per µL

Four studies evaluated participants with CD4 > 100 cells per µL, (Nakiyingi 2014; Pandie 2016; Peter 2015; Peter 2016). Sensitivity estimates ranged from 12% to 19% and specificity estimates ranged from 92% to 100%. See Figure 20. In the five studies that evaluated participants with CD4 ≤ 100 cells per µL, sensitivity estimates ranged from 30% to 65% and specificity estimates ranged from 75% to 94% (Nakiyingi 2014; Pandie 2016; Peter 2012; Peter 2015; Peter 2016). One study (Pandie 2016) had no estimable specificity, as they reported zero TN. The pooled sensitivity (95% CrI) was higher among participants with CD4 ≤100 cells per µL at 54% (38% to 69%) (1239 participants; 41% with TB) versus 17% (10% to 27%), (1519 participants; 28% with TB) among those with CD4 > 100 cells per µL. The pooled specificity was 88% (77% to 94%) for participants with CD4 ≤ 100 cells per µL and 95% (89% to 98%) for those with CD4 > 100 cells per µL. When we limited the analysis to studies involving inpatients with CD4 ≤ 100 cells per µL (Peter 2012; Peter 2016), the pooled sensitivity and specificity were 61% (40% to 78%) and 81% (61% to 91%). Pandie 2016 reported a sensitivity of 50% (95%CI, 21% to 79%) among inpatients with CD4 ≤ 100 cells per µL, but specificity was not estimable and therefore not included in the meta-analysis. Only one study reported data for outpatients with CD4 ≤ 100 cells per µL Peter 2015.

We observed that AlereLAM pooled sensitivity increased as the degree of immunodeficiency increased, from 16% (8% to 31%) in patients with CD4 cell count >200 cells per µL to 24% (14% to 38%) in patients with CD4 count between 101-199 to 54% (38% to 69%) in patients with CD4 ≤ 100. See Figure 21. Also, we observed that a majority of participants contributing data for the CD4 ≤ 200 cells per µL stratum (1825 participants) were participants with CD4 ≤ 100 cells per µL (1239 participants). See Table 23.

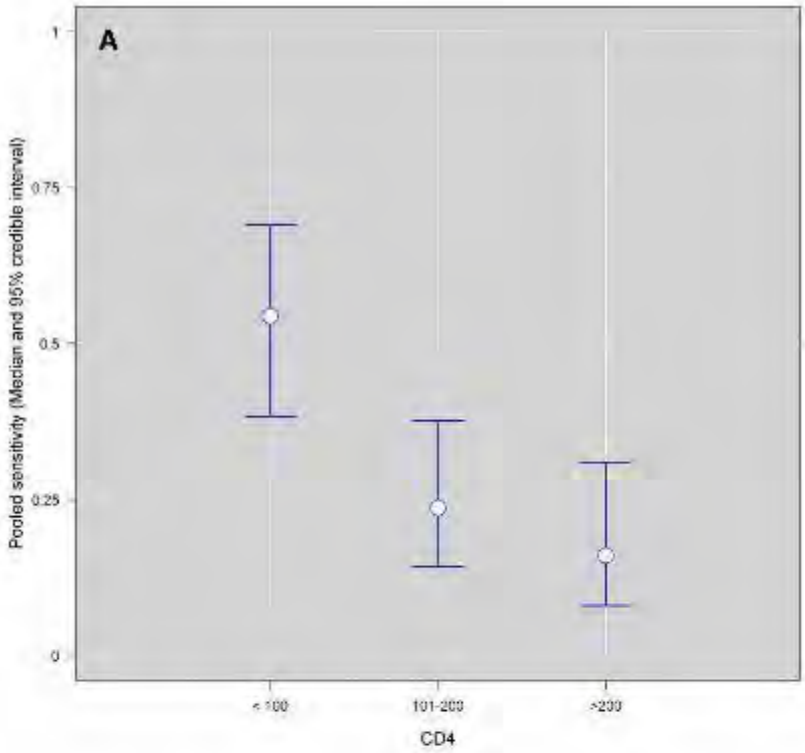
TB prevalence

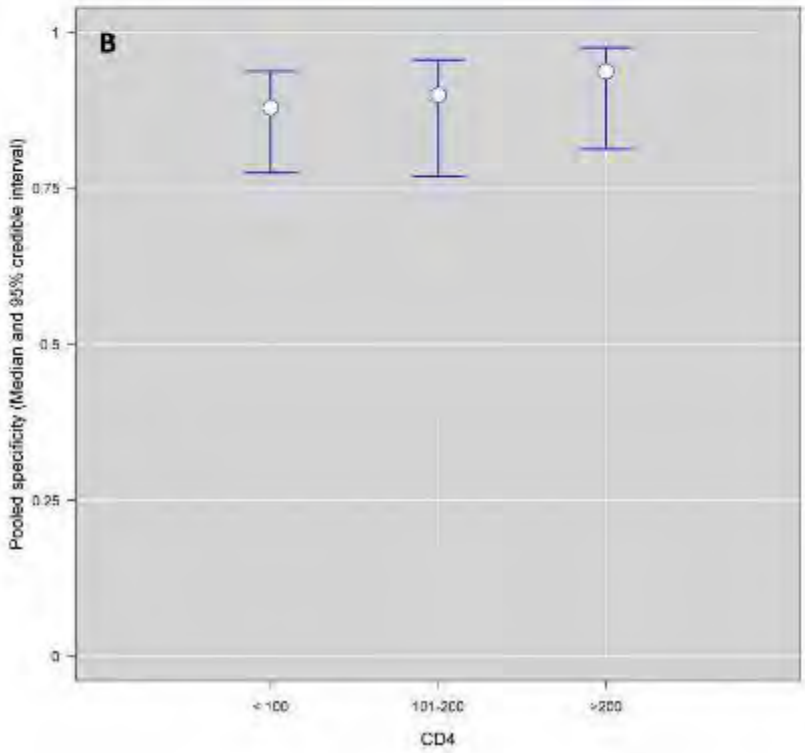
The median prevalence of TB in studies with symptomatic participants was 43% (IQR 32% to 60%). In secondary analysis by TB prevalence, we found that pooled sensitivity and specificity for symptomatic participants in settings with TB prevalence of greater than 43% were 44% (27% to 62%) and 86% (73% to 94%) and 39% (21% to 63%) and 95% (89% to 97%) in settings with a TB prevalence less than 43%. We note that no studies had a TB prevalence of less than 29%.

Sensitivity analyses

When we included all studies with more than 80% symptomatic participants, two studies were re-assigned from 'studies of unselected adults' to 'studies of symptomatic participants' Bjerrum 2015; Lawn 2017. In comparison with estimates without re-classification, pooled sensitivity remained at 42% (33% to 52%) and specificity changed to 93% (88% to 96%), (10 studies, 4331 participants). When we limited the studies to those with low risk of bias for patient selection pooled sensitivity increased to 48% (29% to 67%) and specificity dropped to 82% (61% to 92%), (Peter 2012; Peter 2016, 1413 participants). We did not have enough studies to do sensitivity analysis including only studies with low risk of bias in the reference standard domain. Limiting studies to those that used fresh urine samples (four studies) rather than stored urine sample increased sensitivity to 52% (38% to 68%) with specificity remaining at 91%.

Figure 21. Plot by CD4 of diagnostic accuracy in adults with signs and symptoms.
(A) Sensitivity by CD4 strata; (B) Specificity by CD4 strata. The circle represents the pooled estimates (median), with bars representing 95% credible intervals.





PICO 2: What is the diagnostic accuracy of AlereLAM for TB diagnosis in HIV-positive adults irrespective of signs and symptoms for TB?
Of the 15 studies included, seven studies evaluated the accuracy of AlereLAM for diagnosis in participants irrespective of sign and symptoms ('unselected participants'). The TB prevalence varied from 1% to 33%. Six of the studies reported the proportion of symptomatic participants that were included (e.g. having a positive WHO symptoms screen) which varied from 19% (LaCourse 2016) to more than 90% of participants in two studies (Bjerrum 2015; Lawn 2017). Four studies were carried out in an outpatient setting, one study exclusively in an inpatient setting and two studies in both inpatient and outpatient settings. The median CD4 cell count across studies of unselected adults ranged from 111 to 437 cells per µL across studies. See Appendix 8. Characteristics of Included Studies.

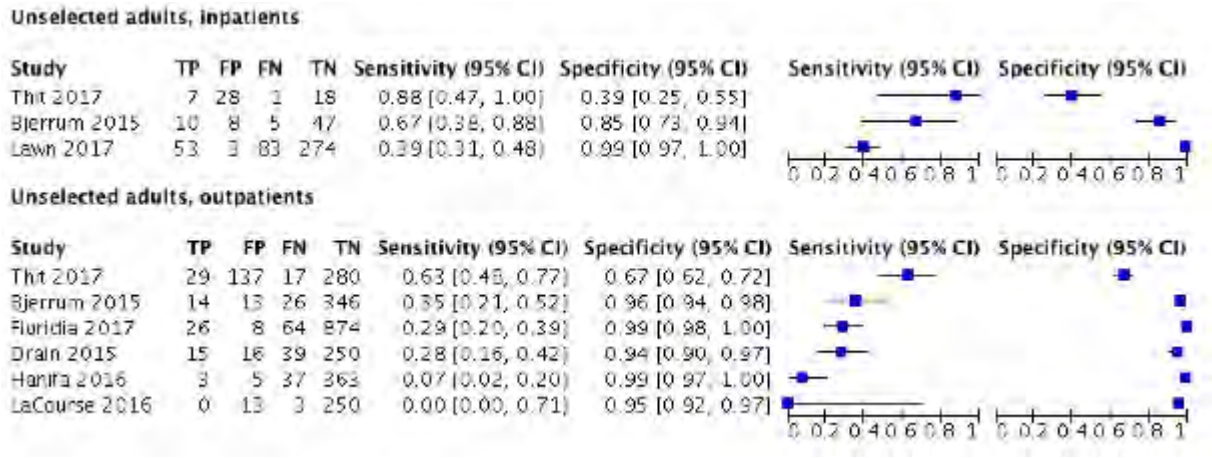
PICO 2a: Accuracy, in inpatient settings
We identified three studies that included inpatients involving 537 participants, 159 (30%) with TB (Bjerrum 2015; Lawn 2017; Thit 2017). Sensitivity estimates ranged from 39% to 88% and specificity estimates ranged from 39% to 99%. The pooled sensitivity and specificity (95% CrI) among inpatients were 62% (41% to 83%) and 84% (48% to 96%). See Figure 22. Thit 2017 reported a very low specificity (39%) and a high sensitivity (88%), based on a relatively small sample size of 54 inpatients; eight (15%) with TB. They reported that 41 of the inpatients (76%) were symptomatic at enrolment with a median CD4 of 96 (IQR 37-277) cells per µL, which is comparable to evaluation of AlereLAM in a population with advanced HIV disease. The study differed from other studies for reasons listed under PICO 2c.

Pooled estimates were not calculated in the original review (Shah 2016) for studies with unselected inpatients due to lack of data.

PICO 2b: Accuracy, in outpatient settings
Outpatients
Six studies were conducted among outpatients, involving 2828 participants; 273 (10%) with TB (Bjerrum 2015; Drain 2015; Floridia 2017; Hanifa 2016; LaCourse 2016; Thit 2017). Sensitivity estimates ranged from 0% to 63% and specificity estimates ranged from 67% to 99%. Pooled sensitivity and specificity (95% CrI) were 31% (18% to 47%) and 95% (87% to 99%). See Figure 22

Pooled estimates were not calculated in the original review (Shah 2016) for outpatients due to lack of data.

Figure 22. Diagnostic accuracy in adults irrespective of signs and symptoms, by setting
Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, by clinical setting.



Type of analysis	Unselected participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
Inpatient	3 studies (537)	159 (30%)	62% (41 to 83)	84% (48 to 96)
Outpatient	6 studies (2828)	273 (10%)	31% (18 to 47)	95% (87 to 99)

Footnote: The proportion of symptomatic participants included ranged from 19% in LaCourse 2016 to 91% in Bjerrum 2015 and in Lawn 2017. See Appendix 8. Characteristics of Included Studies.

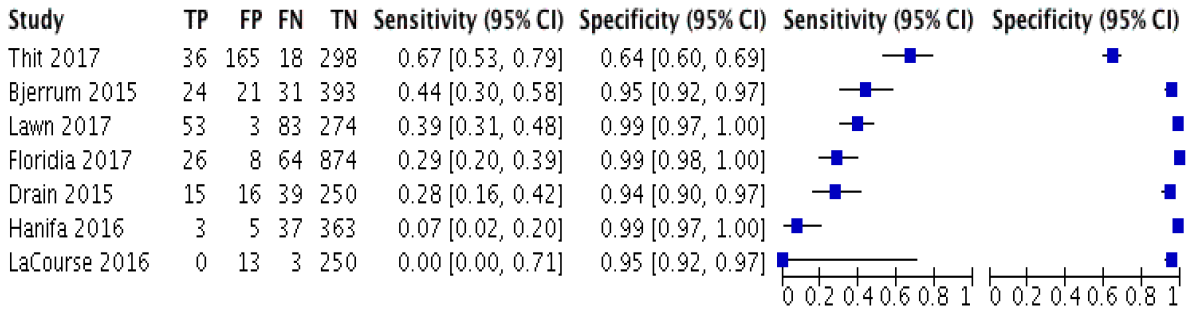
PICO 2c: Overall accuracy, all settings

For the analysis of the overall accuracy of AlereLAM in HIV-positive adults irrespective of signs and symptoms of TB seven studies provided data for 3365 participants; 439 (13%) with TB (Bjerrum 2015; Drain 2015; Floridia 2017; Hanifa 2016; LaCourse 2016; Lawn 2017; Thit 2017). The pooled sensitivity and specificity (95% CrI) were 35% (22% to 50%) and 95% (89% to 98%). Sensitivity estimates ranged from 0% to 67%, and specificity estimates from 64% to 99%. See Figure 23. Sensitivity was lowest (0%) in LaCourse 2016, that differed from the other studies by including a) a population of exclusively pregnant women attending an antenatal care setting, b) a low proportion of symptomatic participants (19%), c) a low TB prevalence (1%), and d) a high median CD4 cell count (437 cells per µL). Specificity was lowest (64%) for Thit 2017 that also reported the highest sensitivity (67%). This study used the new reference scale card with 4 bands and reported that more than 90% of the FP results were grade 1 positive results (classified as positive according to current manufacturer recommendations). Participants included had a median CD4 at 270 cells per mm^3 and 33% were symptomatic at enrolment. The study evaluated sputum samples only and allowed a follow-up for 6 months from AlereLAM testing at enrolment to final classification of participants as 'TB' or 'Not TB' cases. Thit 2017 differed from the other studies by being conducted in Myanmar, and is the only study included in this review that evaluated AlereLAM in a setting outside sub-Saharan Africa.

The pooled sensitivity and specificity of the original review were 30% (20% to 43%) and 94% (86% to 97%) based on three studies and 1055 participants (reported at grade 1 on the old reference scale card with five bands) (Shah 2016).

Figure 23. Diagnostic accuracy in adults irrespective of signs and symptoms, all settings

Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, overall.



Type of analysis	Unselected participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
Overall accuracy	7 studies (3365)	432 (13%)	35% (22 to 50)	95% (89 to 98)

Footnote: The proportion of symptomatic participants included ranged from 19% in LaCourse 2016 to 91% in Bjerrum 2015 and in Lawn 2017. See Appendix 8. Characteristics of Included Studies.

PICO 3: What is the diagnostic accuracy of AlereLAM for diagnosis of TB in adults with advanced HIV disease irrespective of signs and symptoms of TB?

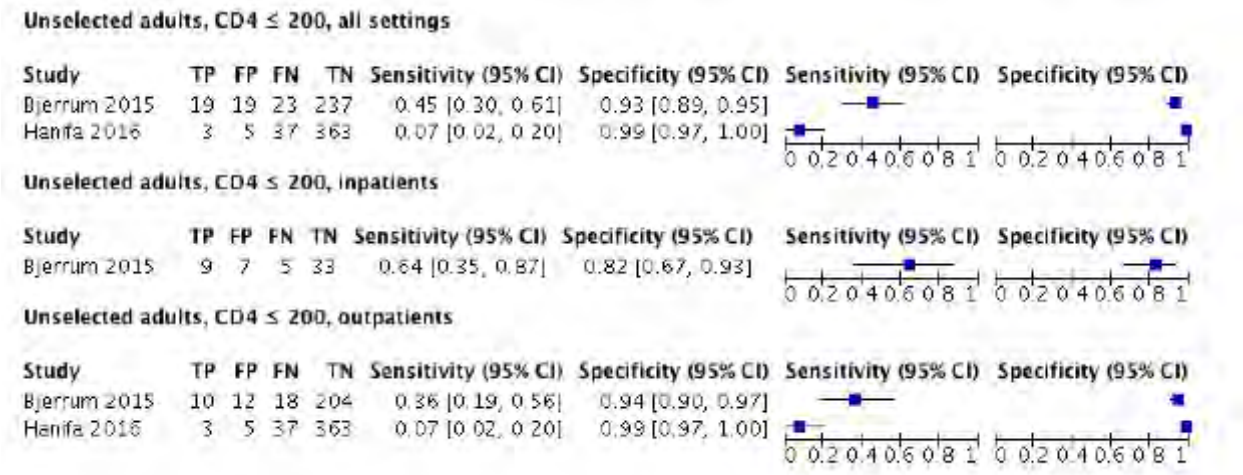
There were limited data to evaluate AlereLAM by CD4 threshold for unselected participants irrespective of signs and symptoms of TB.

PICO 3a: Accuracy, in inpatient setting and CD4 ≤200
For inpatients with CD4 ≤ 200 cells per µL, only one study contributed data and found a sensitivity (95% CI) of 64% (35% to 87%) and specificity of 82% (67% to 93%) (54 participants; 26% with TB) (Bjerrum 2015). See Figure 24.

PICO 3b: Accuracy, in outpatient settings and CD4 ≤200
For outpatients with a CD4≤ 200 cells per µL, two studies contributed data (652 participants; 10% with TB. Sensitivity and specificity were 36% and 94% for Bjerrum 2015, and 7% and 99% for Hanifa 2016. Pooled sensitivity and specificity were 21% (8% to 48%) and 96% (89% to 99%). See Figure 24.

PICO 3c: Accuracy, in all settings and CD4 ≤ 200
Two studies evaluated AlereLAM in unselected participants with CD4 ≤ 200 cells per µL all settings. Sensitivity and specificity were 45% and 93% for Bjerrum 2015, and 7% and 99% for Hanifa 2016. Pooled sensitivity and specificity were 26% (9% to 56%) and 96% (87% to 98%) (706 participants; 12% with TB). See Figure 24.

Figure 24. Diagnostic accuracy in adults irrespective of signs and symptoms, ≤ 200, by setting
Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, CD4 ≤ 200, by setting.



Type of analysis	Unselected participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
CD4 ≤ 200 All settings	2 studies (706)	82 (12%)	26% (9 to 56)	96% (87 to 98)
CD4 ≤ 200 Inpatients	1 study[c] (54)	14 (26%)	Not applicable	Not applicable
CD4 ≤ 200 Outpatients	2 studies (652)	68 (10%)	21% (8 to 48)	96% (89 to 99)

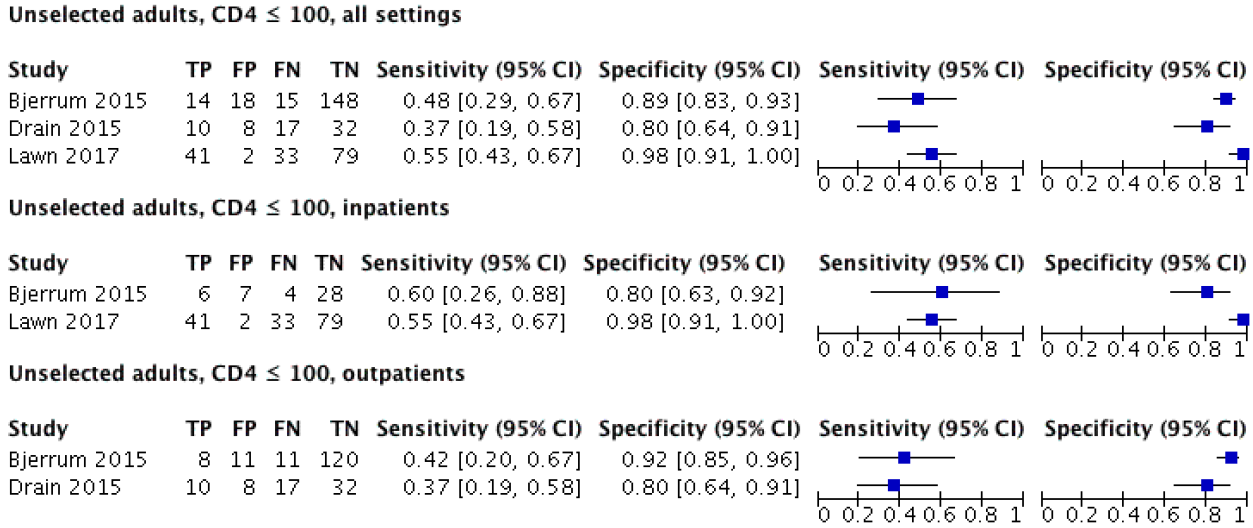
Footnote: The proportion of symptomatic participants included was 91% in Bjerrum 2015 and 53% in Hanifa 2016. See Appendix 8. Characteristics of Included Studies.
[c] Bjerrum 2015, Sensitivity 64% (35% to 87%); Specificity 82% (67% to 93%).

PICO 3d: Accuracy, in inpatient setting and CD4 ≤100
For inpatients with CD4 ≤ 100 cells per µL, two studies contributed data, (200 participants; 42% with TB) (Bjerrum 2015; Lawn 2017). Sensitivity and specificity were 60% and 80% for Bjerrum 2015, and 55% and 98% for Lawn 2017. The pooled sensitivity and specificity were 57% (33% to 79%) and 90% (69% to 97%). See Figure 25.

PICO 3e: Accuracy, in outpatient settings and CD4 ≤100
Two studies evaluated outpatients with CD4 ≤ 100 cells per µL, the pooled sensitivity and specificity were 40% (20% to 64%) and 87% (68% to 94%) (217 participants; 21% with TB) (Bjerrum 2015; Drain 2015). See Figure 25.

PICO 3f: Accuracy, in all settings and CD4 ≤100
Three studies evaluated patients with CD4 ≤ 100 cells per µL, all settings and sensitivity estimates ranged from 37% to 55% and specificity from 80% to 98%. See Figure 25. The pooled sensitivity and specificity (95% CrI) among participants with CD4 ≤ 100 cells per µL were 47% (30% to 64%) and 90% (77% to 96%) for participants with CD4 ≤ 100 cells per µL (417 participants; 31% with TB) (Bjerrum 2015; Drain 2015; Lawn 2017).

Figure 25. Diagnostic accuracy in adults irrespective of signs and symptoms, CD4 ≤ 100, by setting. *Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, CD4 ≤ 100, by setting.*



Type of analysis	Unselected participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
CD4 ≤ 100 All settings	3 studies (417)	130 (31%)	47% (30 to 64)	90% (77 to 96)
CD4 ≤ 100 Inpatients	2 studies (200)	84 (42%)	57% (33 to 79)	90% (69 to 97)
CD4 ≤ 100 Outpatients	2 studies (217)	46 (21%)	40% (20 to 64)	87% (68 to 94)

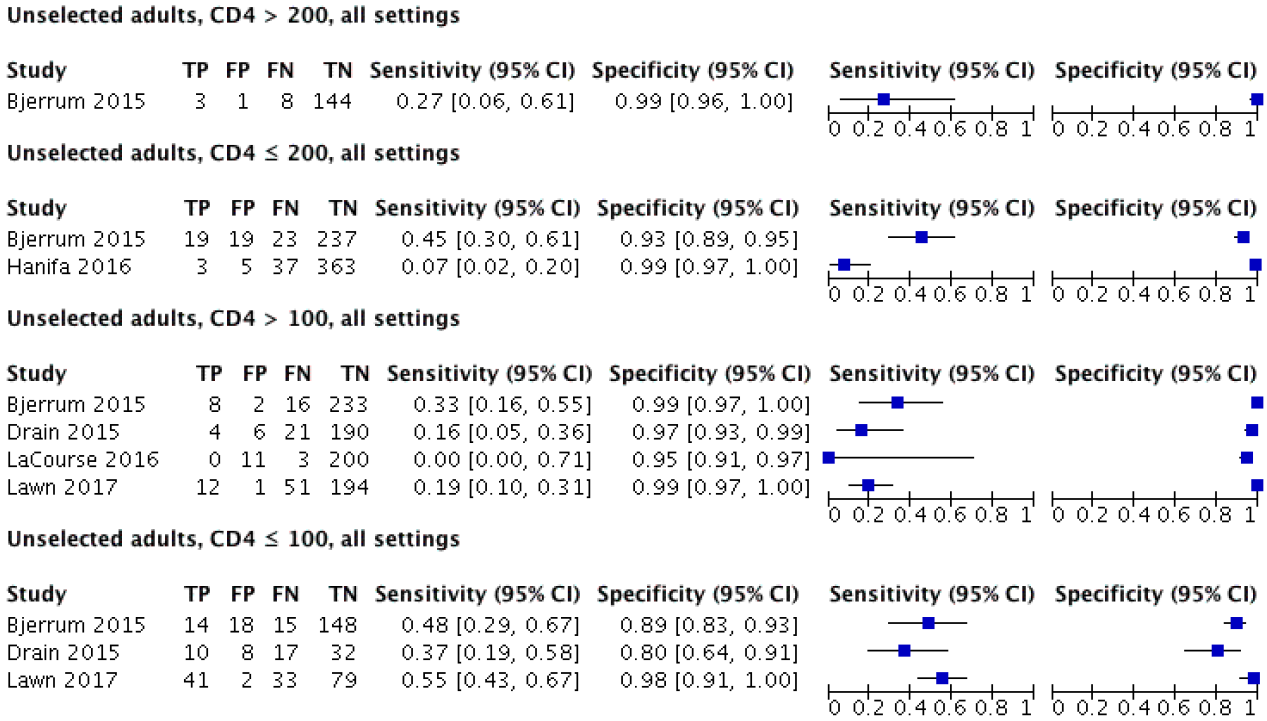
Footnote: The proportion of symptomatic participants included was not stated for Drain 2015 and was 91% in both Bjerrum 2015 and Lawn 2017. See Appendix 8. Characteristics of Included Studies.

Additional investigations of heterogeneity

CD4 count

For comparison to studies evaluating diagnostic accuracy at lower CD4 counts, we assessed diagnostic accuracy among participants with CD4 > 200 cells per µL and CD4 > 100 cells per µL. Only one study reported data for participants with CD4 > 200 cells per µL and reported a sensitivity of 27% (95% CI 6% to 61%) and specificity of 99% (95% CI; 96% to 100%) (Bjerrum 2015). Four studies evaluated AlereLAM in participants with CD4 > 100 cells per µL where sensitivity estimates ranged from 0% to 33% and specificity estimates ranged from 95% to 99%. Pooled sensitivity and specificity (95% CrI) were 20% (10% to 35%) and 98% (95% to 99%), (952 participants, 12% with TB). See Figure 26.

Figure 26. Diagnostic accuracy in adults irrespective of signs and symptoms, by CD4 count.
Forest plots and meta-analysis of AlereLAM sensitivity and specificity for TB against a microbiological reference standard, by CD4.



Type of analysis	Unselected participants			
	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
CD4 > 200	1 study[a] (156)	11 (7%)	Not applicable	Not applicable
CD4 ≤ 200	2 studies (706)	82 (12%)	26% (9 to 56)	96% (87 to 98)
CD4 > 100	4 studies (952)	115 (12%)	20% (10 to 35)	98% (95 to 99)
CD4 ≤ 100	3 studies (417)	130 (31%)	47% (40 to 64)	90% (77 to 96)

Footnote: The proportion of symptomatic participants included ranged from 19% in LaCourse 2016 to 91% in both Bjerrum 2015 and Lawn 2017. See Appendix 8. Characteristics of Included Studies.
[a] Bjerrum 2015, Sensitivity 27% (6% to 61%); Specificity 99% (96% to 100%).

TB prevalence

The median prevalence of TB in studies with unselected participants was 10% (IQR 9% to 17%). In a secondary analysis by TB prevalence, we found that pooled sensitivity and specificity for unselected participants in settings with TB prevalence of 10% or more were 45% (31% to 61%) and 92% (79% to 97%) (4 studies) compared to 16% (5% to 36%) and 98% (94% to 99%) in settings with TB prevalence less than 10% (three studies). In general, TB prevalence increased in studies with a higher proportion of symptomatic participants.

Sensitivity analyses

When we reclassified studies with more than 80% of participants being symptomatic at inclusion as 'studies of symptomatic adults' (Bjerrum 2015; Lawn 2017) pooled sensitivity and specificity changed slightly to 31% (16% to 50%) and 95% (84% to 98%) (five studies, 2483 participants).

Limiting analysis to studies with low risk of bias for patient selection pooled sensitivity increased to 39% (17% to 66%) and specificity dropped to 93% (61% to 92%) (three studies,1338 participants).

Limiting analysis to studies with a low risk of bias in the reference standard domain (two studies), increased pooled specificity to 99% (95% to 99%) while pooled sensitivity decreased to 24% (8% to 53%). As for studies with symptomatic individuals, sensitivity increased in studies evaluating AlereLAM on fresh urine rather than stored urine sample with sensitivity at 41% and specificity at 93% (five studies).

Additional analyses

Investigations of heterogeneity, all studies combined

As we found a similar pattern among studies with symptomatic and studies with unselected participants in regard to performance when stratified by clinical setting and CD4, we investigated heterogeneity for these variables across all 15 studies combined. We present pooled sensitivity and specificity of AlereLAM for all studies combined stratified by setting and by CD4 cell count in

Table 24.

Setting
We identified a total of nine studies evaluating AlereLAM among inpatients and 10 studies among outpatients. The pooled sensitivity (95% CrI) among inpatients was 54% (44% to 67%) (2790 participants) versus 30% (21% to 41%) (3772 participants) among outpatients. Pooled specificity among inpatients was lower at 87% (75% to 94%) versus 95% (86% to 98%) among outpatients.

CD4 cell count
In the combined analysis of participants with CD4 ≤ 100 cells per μL, pooled sensitivity (95% CrI) was 52% (41% to 63%) (7 studies, 1656 participants) versus 18% (12% to 25%) (eight studies, 2471 participants) among those with CD4 > 100 cells per μL. The pooled specificity was 89% (82% to 94%) for participants with CD4 ≤ 100 cells per μL and 97% (94% to 98%) for those with CD4 > 100 cells per μL.

In the combined analysis of participants with CD4 ≤ 200 cells per μL, pooled sensitivity (95% CrI) was 39% (25% to 54%) (six studies, 2531 participants) versus 17% (9% to 30%) (four studies, 894 participants) among those with CD4 > 200 cells per μL. The pooled specificity was 92% (85% to 96%) for participants with CD4 ≤ 200 cells per μL and 96% (87% to 98%) for those with CD4 > 200 cells per μL. When stratified by both setting and CD4 for all studies combined the sensitivity remained higher (and specificity lower among inpatients compared to outpatient in all CD4 strata.

Table 24. AlereLAM pooled sensitivity and specificity for TB diagnosis, combined analysis of studies among symptomatic and unselected participants.

Type of analysis	Studies (total participants)	Participants with TB (%)	Pooled sensitivity (95% CrI)	Pooled specificity (95% CrI)
By setting				
Inpatient	9 studies (2790)	1027 (37%)	54% (44 to 67)	87% (74 to 94)
Outpatient	10 studies (4024)	682 (17%)	30% (20 to 41)	96% (92 to 98)
By CD4 cell				
CD4 > 200	4 studies (894)	174 (19%)	17% (9 to 30)	96% (87 to 98)
CD4 ≤ 200	6 studies (2531)	804 (32%)	39% (25 to 54)	92% (85 to 96)
CD4 > 100	8 studies (2471)	540 (22%)	18% (12 to 25)	97% (94 to 98)
CD4 ≤ 100	7 studies (1656)	642 (39%)	52% (41 to 63)	89% (82 to 94)
By CD4 and setting				
CD4 ≤ 200 inpatients	3 studies (1063)	362 (34%)	56% (39 to 72)	81% (66 to 90)
CD4 ≤ 100 inpatients	4 studies (934)	354 (38%)	60% (46 to 72)	86% (72 to 93)
CD4 101-200 inpatients	3 studies (284)	82 (29%)	37% (20 to 62)	83% (63 to 93)
CD4 > 200 inpatients	2 studies (324)	75 (23%)	21% (9 to 42)	89% (72 to 96)
CD4 > 100 inpatients	4 studies (789)	197 (25%)	23% (4 to 789)	97% (89 to 99)
CD4 ≤ 200 outpatients	3 studies (901)	165 (18%)	22% (11 to 40)	96% (90 to 98)
CD4 ≤ 100 outpatients	3 studies (338)	92 (27%)	36% (21 to 55)	89% (77 to 95)
CD4 101-200 Outpatients	2 studies (222)	60 (27%)	20% (8 to 43)	96% (88 to 99)
CD4 > 200 outpatients	1 study (147)	0 (0%)	Not applicable	Not applicable
CD4 > 100 patients	4 studies (1120)	234 (21%)	19% (11 to 31)	97% (94 to 99)

Abbreviations: Crl: credible interval; AlereLAM: Alere Determine™ TB lipoarabinomannan assay; TB: tuberculosis.

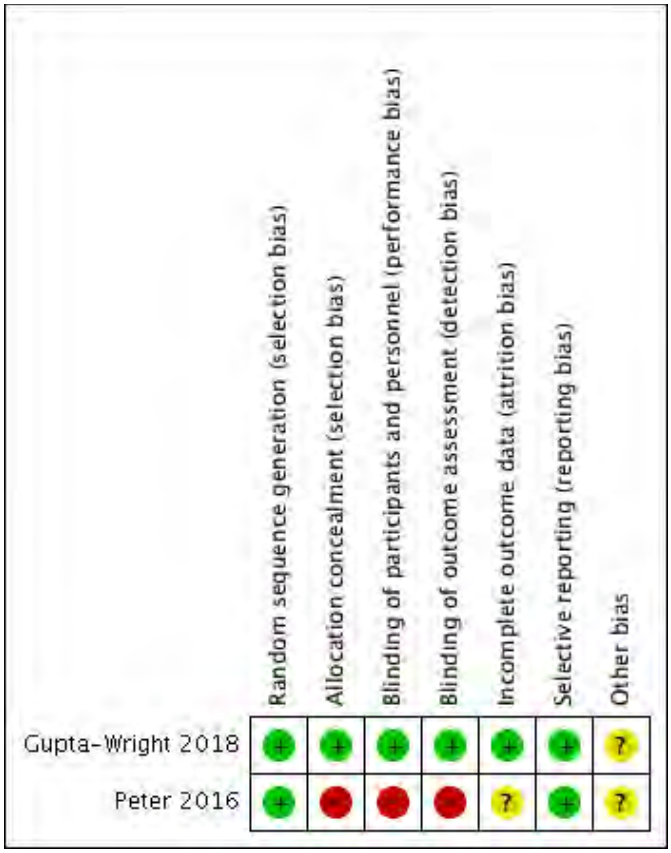
PICO 4: Can the use of AlereLAM in HIV-positive adults reduce mortality associated with advanced HIV disease?

For PICO 4, data were available for the inpatient setting only. We therefore report results for PICO 4b) Inpatient setting; PICO 4e) Inpatient setting with CD4 ≤ 200; and PICO 4h) Inpatient setting with CD4 ≤ 100.

We identified two studies that assessed the impact of AlereLAM on mortality when the test was used for clinical decision-making (Peter 2016; Gupta-Wright 2018a) Both studies were multi-site randomized controlled trials that evaluated the impact of using AlereLAM as a TB diagnostic test to guide treatment initiation in HIV-positive adult inpatients, comparing all-cause mortality at 56 days between the AlereLAM intervention arm and standard-of-care control arm.

Figure 27 *shows the risk of bias assessment for the two studies.*

Figure 27. Risk of bias summary: review authors' judgements about each risk of bias item for each included study.



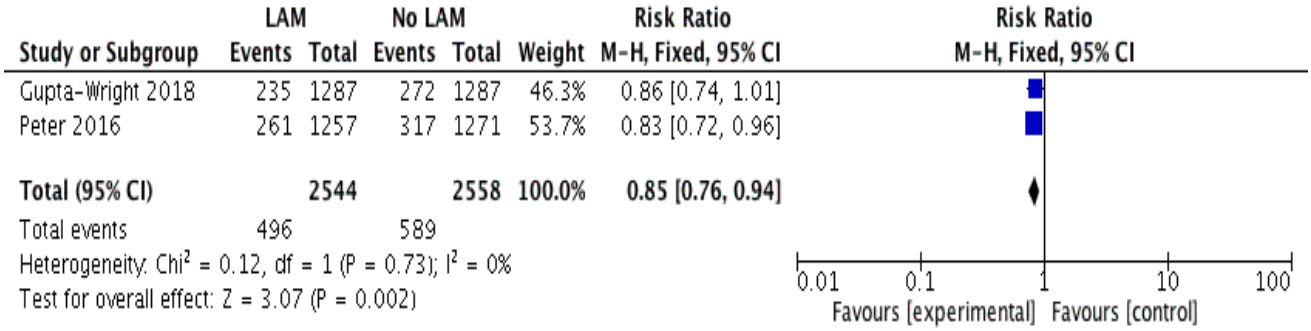
PICO 4b: Impact of AlereLAM on mortality in inpatient settings
Both Peter 2016 and Gupta-Wright 2018a demonstrated that the use of AlereLAM was associated with reduced eight-week mortality, although in Gupta-Wright 2018a, this was only demonstrated in three subgroups (patients with presumed TB, patients with CD4 counts less than 100 cells per μL, and patients with severe anaemia) rather than the overall trial cohort (Table 3 and Table 4).

Peter 2016 found that, in randomly assigned HIV-positive inpatients, AlereLAM in combination with routine diagnostic tests (smear microscopy, Xpert MTB/RIF, and culture) to guide the rapid initiation of TB treatment in HIV-positive adults with at least one TB symptom and illness severity that warranted admission to hospitals in South Africa, Tanzania, Zambia, and Zimbabwe, was associated with a relative risk reduction of 17% (95% CI 4% to 28%) in eight-week mortality compared with routine diagnostic tests alone (no AlereLAM) (Table 3 and Table 4).

Gupta-Wright 2018a randomly assigned HIV-positive inpatients from two hospitals in Malawi and South Africa, to either the standard of care (sputum Xpert MTB/RIF, with the option of sending additional samples for routine TB investigations such as smear microscopy or culture) or intervention (which included urine testing for AlereLAM and Xpert MTB/RIF in addition to sputum Xpert MTB/RIF) irrespective of clinical presentation or TB status. Mortality at 56 days was 21% in the standard-of-care group versus 18% in the intervention group, [adjusted risk reduction (aRD) -2.8% (95% CI -5.8 to 0.3), P = 0.074]. However, in three of the twelve prespecified, but underpowered, subgroups, mortality was lower in the intervention group than in the standard-of-care group for CD4 counts less than 100 cells per µL [aRD -7.1% (95% CI -13.7 to -0.4), P = 0.036]; severe anaemia [-9.0% (95% CI -16.6 to -1.3), P = 0.021]; and patients with clinically suspected TB [aRD -5.7% (95% CI -10.9 to -0.5), P = 0.033] (Table 3 and Table 4).

In the meta-analysis involving both trials, the pooled risk ratio was 0.85 (95% CI 0.76 to 0.94) i.e. study participants undergoing AlereLAM testing had 0.85 times the risk or 15% lower risk of mortality than participants undergoing routine TB diagnostic testing without AlereLAM (Figure 15). The absolute effect was 35 fewer deaths per 1,000 (from 14 fewer to 55 fewer) (high-certainty evidence).

Figure 28. Impact of AlereLAM on mortality in HIV-positive adult inpatients.
Forest plots and meta-analysis of the impact of AlereLAM on mortality, compared to the control study arms that did not include AlereLAM testing.



PICO 4e: Impact of AlereLAM on mortality in inpatients with CD4 ≤200
Peter 2016, reported that in their trial of HIV-positive adult inpatients with at least one TB symptom with CD4 count of ≤ 200 cells per µL (1725 patients), the use of AlereLAM testing (intervention) was associated with a HR of 0.87 (0.72 to 1.04) for mortality (i.e. 13% reduction in mortality) compared to the study arm without AlereLAM testing (Table 3 and Table 4). Gupta-Wright 2018a found that in their trial of unselected HIV-positive adult inpatients, rapid urine-based screening (which included AlereLAM) was associated with an adjusted risk difference of -0.1 (-3.3 to -3.1, P = 0.96) in patients with a CD4 count ≥ 100 cells per µL (Table 3 and Table 4).

PICO 4h: Impact of AlereLAM on mortality in inpatients with CD4 ≤100
Peter 2016, reported that in their trial of HIV positive adult inpatients with at least one TB symptom with a CD4 count ≤ 100 cells per µL (1272 patients), the use of AlereLAM testing (intervention) was associated with a HR of 0.88 (0.72 to 1.08) for mortality (i.e. 12% reduction in mortality) compared to the study arm

without AlereLAM testing (Tables 3 and 4). The greatest reduction in mortality (29%) occurred in the 867 patients with a CD4 count ≤ 50 cells per μL (HR 0.71, 0.56 to 0.90). Gupta-Wright 2018a found that in their trial of unselected HIV-positive adult inpatients, rapid urine-based screening (which included AlereLAM) was associated with an adjusted risk difference of -7.1 (-13.7 to -0.4; P = 0.036) in patients with a CD4 count < 100 cells per μL [adjusted odds ratio of 0.72 (0.53 to 0.98)] (Table 3 and Table 4).

Table 3. Comparison of mortality in randomized trials that evaluated a diagnostic or screening intervention using AlereLAM in HIV-positive participants

Study	Population	Illness severity metrics	Design	Time of mortality assessment	Mortality analysis	Mortality in intervention	Mortality in control	Other outcomes assessed
Gupta-Wright 2018a	2574 HIV-positive adults, inpatients (unselected)	Median CD4 222 cells per µL, Karnofsky score 60, BMI 21.7, median haemoglobin 10.4 g/dL	Randomized controlled trial	56 days	aRD -2.8%, 95% CI: 5.8 to 0.3; P = 0.074	18% (235/1287)	21% (272/1287)	Significant mortality reduction in three subgroups - severe anaemia (aRD -9.0%, 95% CI -16.6 to -1.3; P = 0.021); CD4 < 100 cells per µL (aRD -7.1%, 95% CI -13.7 to -0.4; P = 0.036); clinically suspected TB (-5.7% 95% CI -10.9 to -0.5; P = 0.033). More patients in LAM arm were started on treatment (aHR 1.56, 95% CI 1.29 to 1.88; P < 0.0001).
Peter 2016	2528 HIV-positive adults, inpatients (symptoms)	Median CD4 84 cells per µL, Karnofsky score 50, BMI 18.8, median haemoglobin 9.2 g/dL	Randomized controlled trial	8 weeks	ARR 4% (1% to 7%) aRR 0.83, 95% CI 0.73 to 0.96, P = 0.012	20.8% (261/1257)	24.9% (317/1271)	Greatest mortality reduction in those with CD4 < 50 cells per µL (HR 0.71, 0.56 to 0.90). More patients in LAM arm were started on treatment (52% versus 47%; P = 0.024)

Abbreviations: aRD: adjusted risk difference; ARR: absolute risk reduction; aRR: adjusted risk ratio.

Table 4. Effect of using AlereLAM on mortality, stratified by CD4 group

Study	CD4 group	Effect (95% CI)	CD4 group	Effect (95% CI)
Gupta-Wright 2018a	≥ 100	aOR 0.96 (0.74 to1.25)	< 100	aOR 0.72 (0.53 to 0.98)
Peter 2016	> 100	HR 0.71 (0.53 to 0.96)	≤ 100	HR 0.88 (0.72 to 1.08)
Peter 2016	> 200	HR 0.65 (0.44 to 0.97)	≤ 200	HR 0.87 (0.72 to 1.04)

aOR: adjusted odds ratio; CI: confidence interval; HR: hazard ratio

Important differences between the trials evaluating the impact of AlereLAM on mortality
There were several differences between the two trials. The median CD4 count was lower in Peter 2016 compared to Gupta-Wright 2018a (84 cells per µL versus 227 cells per µL). This, in addition to a lower BMI and Karnofsky score, suggests that the population evaluated in the Peter 2016 study may have been sicker. Overall severity of illness was higher in Peter 2016 (mortality 21% in AlereLAM and in 25% in no AlereLAM arms) compared to Gupta-Wright 2018a (mortality 18% in AlereLAM and 21% in no AlereLAM arms). The percentage of patients on antiretroviral therapy was lower in Peter 2016 than in Gupta-Wright 2018a (48% versus 72%). A greater proportion of participants were started on TB treatment (52% versus 21% in Peter 2016 compared to Gupta-Wright 2018a, reflecting different exclusion criteria (clinical suspicion of TB compared with an unselected population irrespective of symptoms).

Impact of AlereLAM on other patient-important outcomes
Although mortality was the primary patient-important outcome of interest for our data extraction, we also recorded data on other patient-important outcomes. Peter 2016 found that the overall percentage of patients started on TB treatment was higher in the intervention group that received AlereLAM (52% versus 47%; P = 0.024), including a higher proportion that was started on days 0-3 (79% versus 69%; P < 0.0001). Gupta-Wright 2018a found that time to diagnosis was marginally shorter in the AlereLAM intervention group compared to the standard-of-care group [median 0 days (IQR 0 to 1) versus 1 day (0 to 6)]. They reported that increases in TB diagnoses in the intervention group that received AlereLAM were not confined to high-risk subgroups, unlike mortality, with an adjusted absolute risk increase of 7.0% (95% CI 4.1 to 10.0) in TB diagnoses in patients with CD4 counts of 100 cells per µL. Time from diagnosis to treatment was short (median of 1 day, IQR 0 to 1) and did not differ between the group that received AlereLAM and the standard-of-care. However, more patients were started on TB treatment during admission in the group that received AlereLAM (268/1287) compared to the standard-of-care group (182/1287) (aHR 1.56, 95% CI 1.29 to 1.88; P < 0.0001).

Association between AlereLAM positivity with mortality
Association between AlereLAM test positivity with mortality in all settings
We additionally identified 12 studies that had data on the association between AlereLAM positivity and mortality (Table 5) as part of diagnostic accuracy studies (in which AlereLAM was not used for clinical decision making). Three studies evaluated the diagnostic accuracy of AlereLAM (without using results for clinical decision making) in inpatients (LaCourse 2018a; Lawn 2017; Manabe 2014) and six studies evaluated the diagnostic accuracy of AlereLAM (without using results for clinical decision making) in outpatients (Balcha 2014; Drain 2015a; Drain 2017; Hanifa 2016; Lawn 2012b; Peter 2015) and three studies evaluated its use in both inpatients and outpatients (Bjerrum 2015; Huerga 2017; Thit 2017). All studies included only HIV-positive participants except one (Drain 2015a), in which the study population consisted of adults with ≥ two TB symptoms for ≥ two weeks being initiated on TB therapy, of whom 93% were HIV-positive. All studies evaluated adults aside from one (LaCourse 2018a) that evaluated children. All were prospective cohorts or nested prospective cohorts within trials or cross-sectional diagnostic accuracy studies (Table 5). The timing of mortality assessment was highly variable and ranged from 56 days to 12 months. The type of mortality analysis also varied although the majority of prospective cohort studies used hazard ratios.

When considering the association of AlereLAM and mortality (not used for clinical decision making), all prospective studies compared patients who had a positive AlereLAM test with those who had a negative AlereLAM test, with some studies providing additional data on the AlereLAM test status stratified by those

with a confirmed diagnosis of TB, those who did not have TB and those with an inconclusive evaluation for TB. Data on patient outcomes were largely restricted to post-hoc analyses. However, all prospective cohort studies aside from one (Thit 2017) demonstrated a significant association between AlereLAM test positivity and mortality, despite considerable variability in the method of TB diagnosis, provision of treatment and length of follow-up. We note that these investigators did not use the results of AlereLAM to guide treatment initiation.

Association between AlereLAM test positivity with mortality in inpatient settings

LaCourse 2018a reported higher mortality at six months (134/100 person years versus 32/100 person years [AHR 4.61 (95% CI 1.63 to 12.96), P = 0.004] in AlereLAM positive than AlereLAM negative hospitalized HIV-positive children who were evaluated for TB irrespective of clinical suspicion. Lawn 2017 reported higher mortality at 90 days [24.5% versus 7.2%, aOR 4.2 (95% CI 1.50 to 11.75)] in unselected HIV-positive adult inpatient. Of note, AlereLAM was performed on frozen urine specimens that were obtained at the time of enrolment. Manabe 2014 reported higher mortality at six months (40% versus 28%, P = 0.016) in AlereLAM positive than AlereLAM negative hospitalized HIV-positive adults with at least one TB symptom (secondary analysis, Nakiyingi 2014). Manabe 2014 additionally reported higher mortality in AlereLAM positive study participants with confirmed TB (39% versus 20%), those with possible TB (22% versus 17%) and those without evidence of TB (49% versus 31%), compared to those in each of these groups respectively that were AlereLAM negative. In contrast, Thit 2017, which was a hospital-based study in Myanmar with inpatients and outpatients that represents the only study with impact data that was conducted outside Africa, reported that AlereLAM had limited potential clinical utility in their cohort. Four out of the six inpatients who died had a positive AlereLAM test but three received anti-TB therapy prior to death and the fourth had cryptococcal meningitis. Bjerrum 2015 did not report mortality data for inpatients and outpatients separately but found that AlereLAM positive participants had a significantly higher probability of death compared to AlereLAM negative in the overall population (49% versus 14%, P < 0.001) and among those with confirmed TB (54% versus 16%, P = 0.002). Bjerrum 2015 reported that among TB participants who received TB treatment, 31% of those who were AlereLAM positive died compared to only 4% of those who were AlereLAM negative. Among TB participants who did not receive treatment at the time of assessment in the study, 100% of those who were AlereLAM positive died compared to 33% of those who were AlereLAM negative. Huerga 2017 also did not report mortality data for inpatients and outpatients separately but found that mortality was higher in AlereLAM positive compared to AlereLAM negative patients (22.8% versus 8.1%, P < 0.0001), although this difference was not statistically significant amongst confirmed TB patients (confirmed TB patients: 22.8% vs 11.1%, P = 0.130). In a post-hoc analysis, Peter 2013 reported that among inpatients, AlereLAM positive TB participants missed by empirical early treatment had lower CD4 counts and higher median illness severity scores, compared to participants who received early treatment based on clinical decision making.

Association between AlereLAM test positivity with mortality in outpatient settings

We identified six studies that presented results on the association of LAM positivity and mortality in the outpatient setting. Balcha 2014 reported higher mortality (20.0% versus 2.7%, P < 0.001) in AlereLAM positive than AlereLAM negative participants. Drain 2015a reported AlereLAM responses over time. They reported that among participants receiving TB therapy, having a positive AlereLAM test at the two-month visit was associated with an adjusted hazard ratio (HR) of 5.58 for mortality (median follow up time of 49 months) compared to participants with a negative AlereLAM test at the two-month visit. Participants with a positive AlereLAM at six months had an adjusted HR of 42.1 for mortality during study follow-up. They found no difference (adjusted HR 0.99, P = 0.99) in mortality comparing baseline AlereLAM results. Drain 2017 reported that HIV- positive ART-naïve adult outpatients with a positive AlereLAM test was associated with an adjusted hazard ratio (HR) of 4.26 for mortality (follow up time of 12 months). Hanifa 2016 reported a higher mortality [14% versus 5%, HR 3.6 (95% CI 1.2 to 10.5), P = 0.04] in AlereLAM positive compared to AlereLAM negative (using grade 1 as the test positivity threshold) HIV-positive adults attending HIV clinics (CD4 ≤ 200 cells per µL) irrespective of symptoms or presentation. Lawn 2012b found that among 23 TB participants who were AlereLAM positive, five people died (22%) compared to zero deaths (0/36) among TB participants who were AlereLAM negative (secondary analysis, Lawn 2012a). Peter 2015

reported mortality of 25% and 11% in AlereLAM positive and AlereLAM negative participants, respectively. In another secondary analysis to Lawn 2012a, Lawn 2013 reported that AlereLAM sensitivity was 100% among TB participants who died compared to 25% among TB participants who were alive at 90 days (P = 0.002).

General observations on AlereLAM positivity, mortality, and CD4 count

LaCourse 2018a and Drain 2015a adjusted their mortality analysis for baseline CD4 count or percentages but did not report specific mortality data stratified by CD4 count. Lawn 2012a and Lawn 2017 found that those testing AlereLAM positive had lower CD4 counts and a higher prevalence and severity of anaemia (P< 0.001) but mortality analyses evaluated AlereLAM positivity and CD4 count separately.

Association between AlereLAM test positivity with mortality in individuals with CD4 ≤ 200 cells per µL

Drain 2017 reported that in HIV positive ART-naïve adult outpatients with CD4 count of < 200 cells per µL, a positive AlereLAM test (grade 2 on the old reference scale card with five band intensities) was associated with an adjusted hazard ratio (HR) of 2.71 (0.95 to 7.71, P = 0.06) for mortality compared to those who had no evidence of TB and a negative AlereLAM result, which rose to 3.61 (1.69 to 7.71, P = 0.0009) when a positive AlereLAM result of grade 3+ or above was analysed (old reference card).

Association between AlereLAM test positivity with mortality in individuals with CD4 ≤ 100 cells per µL

Thit 2017 reported that of the five deaths among 21 inpatients with TB symptoms and a CD4 T-cell count < 100 cells per µL, three (60%) had a positive AlereLAM result but all of these were on TB treatment prior to death. Balcha 2014 reported that among 21 outpatients with positive AlereLAM results who had not received TB diagnosis (neither by bacteriological nor clinical criteria), five died within six months of inclusion. All five whom had positive WHO symptom screens and had baseline CD4 counts < 100 cells per µL; three had started ART within three months of inclusion but none had started TB treatment. Drain 2015a reported that the overall mean urine AlereLAM test grade decreased from 0.7 (+/- 1.3) at baseline to 0.5 (+/- 1.3) at two months to 0.2 (+/- 0.7) at the six months visit and that these results were similar when stratified by CD4 above/below 100 cells per µL. Drain 2017 reported that in HIV-positive ART-naïve adult outpatients with CD4 count ≤ 100 cells per µL, a positive AlereLAM test (grade 2 on the old reference scale card with five band intensities) was associated with an adjusted hazard ratio (HR) of 2.96 (1.01 to 8.70, P = 0.05) for mortality compared to those who had no evidence of TB and a negative AlereLAM result, which rose to 3.04 (1.34 to 6.91, P = 0.008) when a positive AlereLAM result of ≥ 3+ was analysed.

Although these studies did not directly assess the impact of AlereLAM on patient-important outcomes, Lawn 2012b found that patients who had a positive AlereLAM result initiated treatment within eight days, compared to 21 days for those with a negative test result. Peter 2015 demonstrated that LAM positivity was associated with same day treatment initiation, compared to treatment initiation between day 2 to 56 for those with a negative test result.

Table 5. Comparison of mortality in AlereLAM positive and AlereLAM negative participants in diagnostic accuracy studies

Study	Population	Design and timing for mortality analysis	Population for mortality analysis	Mortality in LAM positive	Mortality in LAM negative	Other outcomes assessed
LaCourse 2018a	181 HIV-positive children, inpatients (unselected)	Nested prospective cohort 6 months	137 HIV-positive inpatient children with valid LAM results	134/100 person years aHR 4.61 95% CI: 1.63-12.96; P = 0.004	32/100 person years	Hazard ratio adjusted for CD4 %.
Lawn 2017	427 HIV-positive adults, inpatients (unselected)	Prospective cohort 90 days	136 TB cases	24.5% (13/53) aOR 4.2 95% CI: 1.50-11.75	7.2% (6/83)	LAM+ participants had a lower CD4 count and more severe anaemia (P < 0.001)
Manabe 2014	506 HIV-positive adults with at least 1 TB symptom, inpatients (symptomatic)	Prospective cohort 6 months	351 enrollees 145 TB cases 21 with possible TB 185 with no evidence of TB	40% (54/134) 39% (35/90) 22% (2/9) 49% (17/35) Unadjusted HR for LAM positivity 1.67; P = 0.025	28% (60/217) 20% (11/55) 17% (2/12) 31% (47/150)	
Thit 2017	517 HIV-positive adults, inpatients (unselected)	Prospective cohort 6 months	54 TB cases	11.4% (4/35)	10.5% (2/19)	
Balcha 2014*	757 HIV-positive adults eligible for ART (CD4 < 350 or WHO stage 4), outpatients (unselected)	Prospective cohort 6 months	148 TB cases	20% (7/35)	2.7% (3/113)	
Bjerrum 2015	469 HIV-positive adults eligible for ART (WHO stage 3/4, CD4 < 350) (unselected)	Prospective cross-sectional 6 months	469 enrollees 55 TB cases 39 TB cases starting treatment	49% 22/45 54% (13/24) 32% (5/16) Kaplan-Meier log-rank test P < 0.001	14% (59/424) 16% (5/31) 4% (1/23)	
Drain, 2015a	90 adults with ≥ 2 TB symptoms for ≥ 2 weeks being initiated on TB therapy, no sputum or smear negative, outpatients (symptomatic)	Prospective cohort 2 months and 6 months	90 outpatients 2 months 6 months	50% (4/8), AHR 5.58 95% CI: 1.24-25.2, P = 0.02 AHR 42.1 95% CI: 1.87-9.52, P = 0.02	18.5% (12/65)	Treatment monitoring: the % of LAM+ participants decreased from 32% (baseline) to 16% (2 months). Hazard ratio adjusted for baseline CD4 count.

Drain 2017	796 HIV-positive adults, ART-naïve, outpatients (unselected)	Prospective cohort 12 months	726 HIV-positive ART-naïve outpatients	31.2% (29/93) MHR 4.26 95% CI: 2.65-6.84 (includes LAM grade 1 positive result)	9.5% (60/633) 8.3% (42/504) LAM negative without evidence of TB 14% (18/129). LAM negative with evidence of TB	Mortality hazard ratios stratified by CD4 levels
Hanifa 2016*	586 HIV-positive adults (CD4 < 200) attending HIV clinics, outpatients (unselected)	Nested within prospective cohort 6 months	426 enrollees with evaluable data	14% (4/28) HR 3.6 95% CI: 1.2-10.5 P = 0.04	5% (20/440)	
Huerga 2017*	474 HIV-positive adults with cough or cough plus other TB symptom, inpatients and outpatients (symptomatic)	Prospective cohort 2 months	468 enrollees with vital status data Confirmed TB patients Non-TB patients Patients with inconclusive TB classification	22.8%, aOR 2.7, 95% CI: 1.5-4.9, P = 0.001 22.8% 15.8% 28.1%	8.1% 11.1% 4.0% 9.9%	
Lawn 2012b*	325 HIV-positive adults, ART-naïve, outpatients (unselected)	Prospective cohort 90 days	59 TB cases	21.7% (5/23) (same at 30 days)	0% (0/36) (same at 30 days)	Time to treatment 8 days (LAM positive) versus 24 days (LAM negative)
Peter 2015	583 HIV-positive adults with suspected TB (symptomatic)	Cross-sectional 6 months	583 enrollees 123 TB cases	25% (9/32) 35% (6/17) ARR 14% P = 0.02	11% (40/361) 14% (15/106)	POC LAM (unclear grade) would have increased same day treatment initiation from 24% (21/89) to 44% (39/89), P = 0.004

Abbreviations: AHR: adjusted hazard ratio; aOR: adjusted odds ratio; ARR: absolute risk reduction; ART: antiretroviral therapy; MHR: mortality hazard ratio; TB: tuberculosis.
*Denotes study in which grade 1 (using the old reference card with five band intensities) was used.

Discussion

This updated systematic review summarizes the current literature and includes 15 unique studies on the accuracy of the urine-based lateral flow lipoarabinomannan assay, Alere Determine™ TB LAM Ag, 'AlereLAM', for tuberculosis (TB) in adults with human immunodeficiency virus (HIV) and integrates nine new studies identified since the original WHO and Cochrane Review (WHO Lipoarabinomannan Policy Guidance 2015; Shah 2016). Eight studies used AlereLAM for TB diagnosis in symptomatic adult participants with signs and symptoms suggestive of TB. These studies largely focused on inpatient settings and had high TB prevalence. Seven studies used AlereLAM for diagnosing TB in unselected participants that may or may not have had symptoms suggestive of TB when enrolled in the study. The studies with unselected participant were conducted predominantly in outpatient settings and, compared to studies with exclusively symptomatic participants, had lower TB prevalence and involved patients with higher CD4 counts; the proportion of symptomatic participants in these studies ranged from 19% to 90%. All studies were conducted in low- and middle-income countries with a high TB/HIV burden, and only one study outside sub-Saharan Africa. We identified two randomized controlled trials that evaluated the impact of AlereLAM implementation among hospitalized patients on morality and other outcomes.

A summary of AlereLAM performance in children is reported separately in Appendix 9: Diagnostic accuracy of AlereLAM among HIV-positive children, summary.

Summary of main results

For TB diagnosis in HIV-positive adults presenting with signs and symptoms of TB, the diagnostic accuracy of AlereLAM is:
in inpatient settings, sensitivity 52% and specificity 87% (PICO 1a)
in outpatient settings, sensitivity 29% and specificity 96% (PICO 1b)
in all settings, sensitivity 42% and specificity 91% (PICO 1c)

For TB diagnosis in HIV-positive adults irrespective of signs and symptoms of TB, the diagnostic accuracy of AlereLAM is:
in inpatient settings, sensitivity 62% and specificity 84% (PICO 2a)
in outpatient settings, sensitivity 31% and specificity 95% (PICO 2b)
in all settings, sensitivity 35% and specificity 95% (PICO 2c)

For diagnosis of TB in adults with advanced HIV disease irrespective of signs and symptoms of TB, the diagnostic accuracy of AlereLAM is (limited data available):
in inpatient setting CD4 ≤ 200, sensitivity of 64% and specificity 82% (one study, PICO 3a)
in outpatient setting CD4 ≤ 200, sensitivity 21% and specificity 96% (PICO 3b)
in all settings CD4 ≤ 200, sensitivity 26% and specificity 96% (PICO 3c)
in inpatient setting CD4 ≤ 100, sensitivity 57% and specificity 90% (PICO 3d)
in outpatient setting CD4 ≤ 100, sensitivity 40% and specificity 87% (PICO 3e)
in all settings CD4 ≤ 100, sensitivity 47% and specificity 90% (PICO 3f)

For diagnosis of TB in HIV-positive children, the diagnostic accuracy of AlereLAM is (limited data available)
in all settings, including all children, for individual studies, sensitivity and specificity were 42% and 94% (outpatient setting); 56% and 95% (inpatient setting); and 43% and 80% (both inpatient and outpatient settings)

For use of AlereLAM to reduce mortality associated with advanced HIV disease (two randomized trials) the pooled risk ratio for mortality was 0.85 (0.76 to 0.94) and the absolute effect was 35 fewer deaths per 1,000 (from 14 fewer to 55 fewer) (PICO 4)

AlereLAM for TB diagnosis in symptomatic participants

For TB diagnosis among symptomatic adults, the pooled sensitivity of AlereLAM was 42% and pooled specificity was 91%. In planned investigations of heterogeneity, we found an inverse correlation between AlereLAM sensitivity and CD4 count, with increasing sensitivity as patient CD4 count decreased (increased from 16% in patients with CD4 cell count > 200 cells per µL to 24% in patients with CD4 cell count between 101-199 cells per µL, to 54% in patients with CD4 ≤ 100 cells per µL. Similarly, we a priori planned to investigate and expected to find higher sensitivity in patients who were hospitalized (sensitivity increased from 29% among outpatients to 52% among inpatients) while specificity decreased (from 96% among outpatients to 87% among inpatients).

Results of these studies indicate that in theory, for a population of 1000 people where 300 have microbiologically-confirmed TB, 189 would be AlereLAM-positive: of these, 63 (33%) would not have TB (false-positives); and 811 would be AlereLAM-negative: of these, 174 (21%) would have TB (false-negatives).

AlereLAM for TB diagnosis in unselected participants

For TB diagnosis among unselected HIV-positive adults (with or without signs or symptoms of TB), the pooled sensitivity was low (35%), with a relatively high pooled specificity (95%). In the investigations of heterogeneity, we expected and found a higher sensitivity in patients with low CD4 cell count and among inpatients compared to patients with higher CD4 cell and outpatients respectively, though data to inform subgroup analyses were limited. We noted that participants included in the studies with unselected participants often presented with sign and symptoms suggestive of TB (a positive WHO TB screen), and in the studies evaluating inpatients the majority of participants (> 80%) were in fact presenting with signs and symptoms suggestive of TB. These studies may be considered more similar to studies with exclusively symptomatic participants. In additional analysis of heterogeneity, we examined diagnostic accuracy based on TB prevalence within the studied cohort, as an alternative surrogate to presence of symptoms or CD4 count as an assessment of pre-test probability. We found that pooled sensitivity was 45% when TB prevalence within the study population was ≥ 10%, compared to only 16% when TB prevalence in the study population was < 10%.

Results of these studies indicate that in theory, for a population of 1000 people where 100 have microbiologically-confirmed TB, 80 would be AlereLAM-positive: of these, 45 (56%) would not have TB (false-positives); and 920 would be AlereLAM-negative: of these, 65 (7%) would have TB (false-negatives).

AlereLAM for TB diagnosis, overall

The findings of this updated review are consistent with those of the original review (WHO Lipoarabinomannan Policy Guidance 2015; Shah 2016). Inclusion of additional studies in this updated review provided the basis for a more precise estimate of the AlereLAM overall sensitivity and specificity. It further allowed us to address key questions regarding test accuracy and sources of heterogeneity including clinical setting and CD4 cell count in studies with symptomatic individuals and in studies with unselected participants.

Overall, we found lower sensitivity for diagnosis of TB among people living with HIV than the internationally suggested target of minimum 65% overall for rapid non-sputum TB tests (WHO TTP 2014). We found that sensitivity increased when considering inpatients and individuals with lower CD4 counts, whether considering studies with exclusively symptomatic participants or those with unselected participants.

When restricting analysis to studies that included participants unable to produce a sputum sample, the estimates of sensitivity increased. Sputum-scarce patients may be the potential target population to benefit the most from urine-based testing as they cannot have other sputum-based diagnostic testing and are likely to have high yield of urine LAM test positivity (Sabur 2017). However, only a few studies included patients who could not provide sputum samples for diagnostic testing. To the extent that inability to produce sputum is correlated with severity of TB disease and/or LAM positivity, this approach to participant selection could have lowered sensitivity estimates within these studies.

Sensitivity analysis further revealed a higher sensitivity among studies evaluating AlereLAM on fresh non-stored urine samples without it affecting specificity. However, no study has made a direct comparison of performance on fresh versus frozen/stored urine samples and the significance of this is unclear.

Overall, we found that the estimated specificities were approaching the recommended targets for non-site-specific, non-sputum based test (WHO TTP 2014), although lower specificity was found among inpatients and those with advanced immunosuppression compared to outpatients and those with higher CD4 counts. We expected that, if restricting the analysis to studies using a higher quality reference standard (e.g. inclusion of more than one specimen type), that estimates of specificity would increase, but had limited data to conduct such a sensitivity analysis.

In a diagnostic test accuracy systematic review, the reference standard is the best available test to determine the presence or absence of the target condition. We only included studies with a microbiological reference standard, which is considered the best currently available reference standard for TB. We included studies that evaluated AlereLAM for diagnosis of pulmonary TB, extrapulmonary TB, or both pulmonary and extrapulmonary TB. However, we recognize that a substantial number of TB cases may not be verified by microbiological testing if only sputum is tested and when patients with advanced HIV are assessed. We acknowledge difficulties in diagnosing HIV-associated TB with extrapulmonary and disseminated forms of disease and considered a standardized reference standard using two or more specimen types to be of higher quality than a reference standard using one specimen type. The higher quality reference standard is better at classifying which patients have and do not have TB. A lower quality reference standard may miss some TB cases and classify some TB patients as not having TB. This may make a truly positive AlereLAM result seem like an FP leading to an underestimation of specificity. In this review, we did not assess performance against a composite reference standard that uses microbiological or clinical information to classify TB. This was done in the original WHO and Cochrane Review (WHO Lipoarabinomannan Policy Guidance 2015; Shah 2016), but found little impact on pooled estimates of sensitivity and specificity relative to performance measured against a microbiological reference standard.

We could not determine whether heterogeneity in specificity estimates was fully attributable to misclassification bias. Some studies (Qvist 2014 and Nel 2017) have postulated that infection with (disseminated) non-tuberculous mycobacteria may also result in false-positive results, although this hypothesis is still questioned (Gupta-Wright 2018). Only one study, Thit 2017, was conducted outside of sub-Saharan Africa, and was noted to report the lowest specificity estimates of all included studies; reasons for potential false-positive results remain unclear and it is unknown if differences in the epidemiology of disseminated NTM and other opportunistic infections across settings could contribute to variation in specificity.

We decided a priori to evaluate performance of AlereLAM in HIV-positive individuals with signs and symptoms of TB (symptomatic) separately from HIV-positive individuals irrespective of signs and symptoms of TB (unselected participants). We considered evaluating AlereLAM performance among specifically *asymptomatic* (i.e. exclusively those without symptoms) participants to assess the role of AlereLAM for TB screening, but such data were lacking among included studies. We did find that several studies among unselected participants reported that a high proportion of study participants had signs and symptoms of TB, suggesting relative similarities to studies that enrolled exclusively symptomatic participants. Consequently, the overall performance of AlereLAM among asymptomatic patients remains largely unknown.

The overall differences in pooled estimates of sensitivity and specificity between studies of symptomatic versus unselected participants may have been attributable to differences in study setting and relative degree of immunosuppression of included participants, rather than type of study (i.e. unselected versus symptomatic participants). When examining inpatients, the pooled estimates for sensitivity were 52% (40% to 64%) and 62% (41% to 83%), when comparing studies of symptomatic participants and those including unselected

participants. Among outpatients, the pooled sensitivity was 29% (17% to 47%) compared to 31% (18% to 47%) among studies of symptomatic participants and unselected participants, respectively. In a secondary analysis combining studies among symptomatic participants and unselected participants, we found a pooled sensitivity of 54% for inpatients compared to 30% among outpatients and a pooled sensitivity of 52% versus 18% among participants with CD4 ≤ 100 cells per µL and CD4 > 100 cells per µL, respectively. In the analysis of all studies combined, the sensitivity remained higher for inpatients than for outpatients across all CD4 strata. This indicate that other characteristics than lower CD4 may explain the higher sensitivity among inpatients like higher TB prevalence, higher mycobacterial burden, renal or genitourinary tract TB with LAM secretion in urine.

Overall, our findings suggest that the diagnostic accuracy of AlereLAM may vary by study setting, CD4 count, and TB prevalence among the target population. The authors hypothesize that these attributes (inpatients, low CD4 counts, or high TB prevalence) may collectively be surrogate indicators of participants with advanced TB disease or higher bacillary burden and LAM antigenuria in whom AlereLAM may aid in the diagnosis of TB, including both pulmonary and extrapulmonary TB. Although subgroup comparisons in diagnostic accuracy reviews are observational and suffer from the same limitations as all observational findings (for example, confounding between characteristics), there is a scientific rationale for these finding in that inpatients, those with low CD4, or cohorts with higher TB prevalence are likely to have higher disease severity or higher bacillary burden. While the test does not identify all TB cases, our findings suggest that it may be of particular value in diagnosing TB among patients with increased disease severity. Other factors that may be considered in evaluating AlereLAM may include ability to perform the test on individuals unable to produce sputum who cannot be diagnosed with other TB diagnostic tests, and ability to implement the test at the point-of-care with non-invasive specimen collection (WHO TTP 2014).

Impact on mortality
We sought to systematically analyse data on patient-important outcomes. Since the publication of the original Cochrane Review (Shah 2016), a second randomized trial that evaluated the impact of AlereLAM implementation on mortality in unselected HIV-positive inpatients (i.e. as a screening test rather than diagnostic test used in patients with TB symptoms) has been published (Gupta-Wright 2018).

Both trials demonstrated mortality reduction in patients with a CD4 count < 100 cells per µL. Both trials also demonstrated an increase in the number of patients started on treatment. Importantly, Gupta-Wright 2018 demonstrated that only 57% of patients could produce sputum for Xpert MTB/RIF testing, in contrast to 99% of patients who could produce urine for AlereLAM testing. Of note, in both trials, patients who could not give informed consent were ineligible to participate. This accounted for 1074/9728 (12.3%) in Peter 2016 and 654/4788 (13.7%) in Gupta-Wright 2018. Since some of these patients could not consent due to the severity of their illness, this may have biased the effect of the intervention towards the null. Since both trials were conducted in sub-Saharan Africa, it is possible that this may limit applicability to other populations.

In additional analyses we demonstrated that within diagnostic accuracy studies that included follow-up for clinical outcomes, without using AlereLAM results for clinical decision making, there appeared to be an association between AlereLAM positivity among both participants with and without confirmed TB (by microbiological and/or clinical study reference standards) and mortality. These data must be interpreted cautiously as they represent secondary analyses within observational cohorts, are limited in size, and may not control for important biases or other factors. It is likely that these findings may represent the effect of missed diagnoses (that could be averted through earlier diagnosis using rapid AlereLAM testing) and/or that there is a biological association between disease severity resulting in AlereLAM excretion in urine.

Differences from original Cochrane Review

In comparison to the original Cochrane Review (Shah 2016), this updated review includes 15 published studies (eight among studies with symptomatic participants, seven among studies with unselected participants). By contrast, the prior review included data from twelve studies, of which three used an older

threshold for determining test positivity that is no longer recommended and three were abstracts of which one was included in this review as an updated published manuscript (Lawn 2014a).

In the evaluation of diagnostic accuracy among symptomatic participants, the pooled estimates for sensitivity (42% versus 45%) and specificity (91% versus 92%) remained similar, comparing the current review and prior review, respectively. When stratified by setting, the pooled estimates among inpatients for sensitivity (52% versus 53%) and specificity (87% versus 90%) did not change substantially when comparing the current review and the prior review, respectively. Pooled estimates among outpatients at the current manufacturer threshold for positivity were not previously available.

In the prior review (Shah 2016), some studies were classified as 'TB screening' if they included participants irrespective of symptoms (i.e. with or without symptoms). Recognizing that these studies may have included a large proportion of symptomatic participants, these studies have been more clearly labelled as studies among 'unselected participants' in the current review. In the prior review, there was insufficient data to perform meta-analysis among unselected participants at the currently recommended manufacturer threshold for test positivity. There were previously insufficient data to investigate heterogeneity due to study setting or CD4 count. In the current review, we report on diagnostic accuracy among inpatients and outpatients and by CD4. In both the current and prior review, data to assess diagnostic accuracy among asymptomatic participants (without signs or symptoms of TB) were unavailable.

This updated review did not assess diagnostic accuracy at an older threshold for determining test positivity (grade 1 out of 5, on an older reference card); data on diagnostic accuracy for this threshold can be found in the prior review (Shah 2016). Similarly, in this updated review we did not evaluate accuracy against a composite reference standard, results of which were included in the prior review (Shah 2016). Finally, given relative lack of data on evaluating the incremental yield of AlereLAM in combination with sputum smear microscopy and Xpert MTB/RIF, we did not include these analyses in the updated review, but a summary of available data is found in the prior review (Shah 2016).

We note that the band intensity of grade 1 in this review corresponds to the current manufacturer threshold for positivity (equivalent to that of grade 2 on the old manufacturer reference card) and all results were evaluated against a microbiological reference standard.

Strengths and weaknesses of the review

The findings in this review are based on comprehensive searching, strict inclusion criteria, and standardized data extraction. The strength of our review is that it enabled an assessment of the accuracy of AlereLAM in people living with HIV with signs and symptoms of TB and irrespective of signs and symptoms of TB. This updated review included new studies published since the original review. However, we found considerable heterogeneity across studies with respect to clinical setting, CD4 count and TB prevalence. For some analyses and subgroup analyses, few studies and participants contributed data and results should, therefore, be interpreted with caution.

The review was further limited by the number of studies that used a lower quality reference standard and the high risk of selection bias in several studies due to exclusion of patients unable to produce sputum. Moreover, only a single study was conducted outside sub-Saharan Africa.

We had overall low concern about the applicability of the included studies to our review question as assessed by QUADAS-2. However, studies of HIV-positive adults irrespective of signs and symptoms had low representation of asymptomatic individuals, as a majority of participants in fact presented with signs and symptoms of TB. To date, no study evaluated AlereLAM in an asymptomatic population.

Using the GRADE approach, we judged the evidence for diagnostic accuracy of AlereLAM to be of low or very low certainty. This means that our confidence in the effect estimate is limited and the true effect may be substantially different from the estimate of the effect.

Authors' conclusions

Implications for practice

For people living with HIV, this review found overall lower sensitivity of AlereLAM than the internationally suggested target of minimum 65% overall for non-sputum based TB diagnostic tests (WHO TTP 2014). This was consistent whether the test is used for diagnosis of TB among symptomatic (sensitivity of 42%) or unselected participants (sensitivity of 35%). The estimated sensitivity suggests that if AlereLAM were to be used alone, more than half of all TB cases would be missed.

Despite the estimated sensitivity, two randomized controlled trials implementing AlereLAM in high prevalence settings in sub-Saharan Africa have demonstrated reduced mortality and impact on other clinical outcomes when used to guide TB treatment in hospitalized HIV-positive adults.

The proposed role for the AlereLAM test is to be used in combination with existing TB tests to assist TB diagnosis and possibly improve important outcomes among HIV-positive patients with advanced disease. The test does not require sputum collection and is not site-specific. Other favorable test characteristics include low-cost, rapidity (< one hour), ease of use (does not require extensive sample preparation), and the fact that the test does not require electricity or special instruments and equipment (WHO TTP 2014). As a simple point-of-care test that does not depend upon sputum evaluation, AlereLAM testing may be the only possible way to confirm a diagnosis when a sputum sample cannot be produced.

Findings suggest that sensitivity increases with lower CD4 counts and in inpatient settings compared to outpatient settings. We found differences in AlereLAM performance based on TB prevalence of the target population (when stratified at greater or less than 10% among unselected participants irrespective of symptoms), with higher diagnostic sensitivity when the study population had higher TB prevalence (≥ 10%). The overall association of differing diagnostic accuracy of AlereLAM when examined by study setting and degree of immunosuppression were consistent irrespective of approach to patient selection. However, we had limited data and these findings should be interpreted with caution.

Clinicians must consider the need for additional testing when interpreting negative AlereLAM results. The consequences of false-negative results are increased risk of morbidity and mortality, delayed treatment initiation, and the continued risk of TB transmission. The consequences of false- positive results are delayed alternative diagnosis, likelihood of anxiety and morbidity caused by additional testing, unnecessary treatment, and possible adverse events; possible stigma associated with a diagnosis of TB. As AlereLAM does not offer information about drug resistance, a culture- or molecular-based diagnosis should be attempted to enable drug susceptibility testing to avoid that patients with unidentified drug-resistant TB may be inappropriately treated with a regimen appropriate only for drug-sensitive disease.

Implications for research

Future studies that evaluate the diagnostic accuracy of non-sputum-based tests for TB, such as AlereLAM, in people living with HIV should use a reference standard that includes at least two specimen types or extrapulmonary specimens in addition to sputum. Moreover, future studies should include patients unable to expectorate sputum in the analysis. While some studies enrolled unselected participants, our review suggests that a large proportion were symptomatic, particularly in the inpatient setting. These features of study design may decrease the risk of bias in the accuracy estimates. Performance of AlereLAM for TB detection among a cohort of exclusively asymptomatic participants is largely unknown. The indication of increased sensitivity with use of fresh urine needs further investigations, and studies in settings outside sub-Saharan Africa are lacking. Further research on effective implementation of AlereLAM within routine clinical practice is needed because the test can only influence clinical practice if the results are believed and acted upon.

Acknowledgements
We are grateful to Vittoria Lutje, the CIDG Information Specialist, for help with the search strategy. The CIDG editorial base is funded by UK aid from the UK government for the benefit of low- and middle-income countries (project number 300342-104). The views expressed do not necessarily reflect the UK government's official policies.

We thank all authors of the included studies for answering our questions and providing data.

Contributions of authors
MS and SB reviewed articles for inclusion and extracted data. MS, SB, IS, ND, and KRS analysed the data. MS, SB, IS, ND, CMD, and KRS interpreted the analyses. SB, MS, and KRS drafted the manuscript in adults. ME and KRS drafted the summary in children. RRN reviewed articles with impact data for inclusion, extracted and analysed data, and drafted the sections on impact. MK, SB, MS, and KRS drafted the GRADE tables. ND and IS drafted the statistical analysis section and the statistical appendix. All authors provided critical revisions to the manuscript. All review authors read and approved the final manuscript draft.

Declarations of interest
Development of the systematic review was in part made possible with financial support from the United States Agency for International Development (USAID) administered by the World Health Organization (WHO) Global TB Programme, Switzerland. KRS, MS, ND, and SB received funding to carry out the review from USAID. KRS has also received financial support for the preparation of systematic reviews and educational materials, consultancy fees from FIND (for the preparation of systematic reviews), honoraria, and travel support to attend WHO guideline meetings. CMD is employed by FIND, a Swiss non-profit organization and WHO Collaborating Centre for Diagnosis. FIND provided funding for an initial assessment of data available for the review. The review authors have no financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the review apart from those disclosed.

References

Included studies

Bjerrum 2015
Bjerrum S, Kenu E, Lartey M, Newman MJ, Addo KK, Andersen AB, et al. Diagnostic accuracy of the rapid urine lipoarabinomannan test for pulmonary tuberculosis among HIV-infected adults in Ghana - findings from the DETECT HIV-TB study. BMC Infectious Diseases 2015;15:407.

Drain 2015
Drain PK, Losina E, Coleman SM, Giddy J, Ross D, Katz JN, et al. Value of urine lipoarabinomannan grade and second test for optimising clinic-based screening for HIV-associated pulmonary tuberculosis. Journal of Acquired Immune Deficiency Syndromes 2015;68(3):274-80.

Drain 2016
Drain PK, Gounder L, Sahid F, Moosa MY. Rapid urine LAM testing improves diagnosis of expectorated smear-negative pulmonary tuberculosis in an HIV-endemic region. Scientific Reports 2016;6:19992.

Floridia 2017
Floridia M, Ciccacci F, Andreotti M, Hassane A, Sidumo Z, Magid NA, et al. Tuberculosis case finding with combined rapid point-of-care assays (Xpert MTB/RIF and Determine TB LAM) in HIV-positive individuals starting antiretroviral therapy in Mozambique. Clinical Infectious Diseases 2017;65(11):1878-83.

Hanifa 2016
Hanifa Y, Fielding KL, Chihota VN, Adonis L, Charalambous S, Karstaedt A, et al. Diagnostic accuracy of lateral flow urine LAM assay for TB Screening of adults with advanced immunosuppression attending routine HIV care in South Africa. PLoS One 2016;11(6):e0156866-.

Huerga 2017
Huerga H, Ferlazzo G, Bevilacqua P, Kirubi B, Ardizzoni E, Wanjala S, et al. Incremental yield of including determine-TB LAM assay in diagnostic algorithms for hospitalized and ambulatory HIV-positive patients in Kenya. PLoS One 2017;12(1).

Juma 2017
Juma F O, Yonga G, Waweru P, Revathi G, Nelson M, Shah R. The diagnostic accuracy of Determine TB LAM antigen in detection of extra pulmonary tuberculosis (EPTB). Journal of Infection 2017;75(6):583-6.

LaCourse 2016
LaCourse SM, Cranmer LM, Matemo D, Richardson B, Kinuthia J, John-Stewart G, et al. Active tuberculosis case finding in HIV-infected pregnant women in Kenya reveals poor performance of symptom screening and rapid diagnostic tests. Journal of Acquired Immune Deficiency Syndromes 2016;71(2):219-27.

Lawn 2017
Lawn SD, Kerkhoff AD, Burton R, Schutz C, Boulle A, Vogt M, et al. Diagnostic accuracy, incremental yield and prognostic value of Determine TB-LAM for routine diagnostic testing for tuberculosis in HIV-infected patients requiring acute hospital admission in South Africa: a prospective cohort. BMC Medicine 2017;15(1):67-.

Nakiyingi 2014
Nakiyingi L, Moodley VM, Manabe YC, Nicol MP, Holshouser M, Armstrong DT, et al. Diagnostic accuracy of a rapid urine lipoarabinomannan test for tuberculosis in HIV-infected adults. Journal of Acquired Immune Deficiency Syndromes 2014;66(3):270-9.

Pandie 2016

Pandie S, Peter JG, Kerbelker ZS, Meldau R, Theron G, Govender U, et al. The diagnostic accuracy of pericardial and urinary lipoarabinomannan (LAM) assays in patients with suspected tuberculous pericarditis. Scientific Reports 2016;6:32924-.

Peter 2012
Peter JG, Theron G, van Zyl-Smit R, Haripersad A, Mottay L, Kraus S, et al. Diagnostic accuracy of a urine lipoarabinomannan strip-test for TB detection in HIV-infected hospitalised patients. European Respiratory Journal 2012;40(5):1211-20.

Peter 2015
Peter J, Theron G, Chanda D, Clowes P, Rachow A, Lesosky M, et al. Test characteristics and potential impact of the urine LAM lateral flow assay in HIV-infected outpatients under investigation for TB and able to self-expectorate sputum for diagnostic testing. BMC Infectious Diseases 2015;15:262.

Peter 2016
Peter JG, Zijenah LS, Chanda D, Clowes P, Lesosky M, Gina P, et al. Effect on mortality of point-of-care, urine-based lipoarabinomannan testing to guide tuberculosis treatment initiation in HIV-positive hospital inpatients: a pragmatic, parallel-group, multicountry, open-label, randomised controlled trial. Lancet 2016;387(10024):1187-97.

Thit 2017
Thit SS, Aung NM, Htet ZW, Boyd MA, Saw HA, Anstey NM, et al. The clinical utility of the urine-based lateral flow lipoarabinomannan assay in HIV-infected adults in Myanmar: an observational study. BMC Medicine 2017;15(1):145-.
Additional studies for impact and association with mortality
Drain 2015a
Drain PK, Gounder L, Grobler A, Sahid F, Bassett, IV, Moosa MYS. Urine lipoarabinomannan to monitor antituberculosis therapy response and predict mortality in an HIV-endemic region: a prospective cohort study. BMJ Open 2015;5(4):e006833. [DOI: http://dx.doi.org/10.1136/bmjopen-2014-006833].

Drain 2017
Drain PK, Losina E, Coleman SM, Giddy J, Ross D, Katz JN, et al. Clinic-based urinary lipoarabinomannan as a biomarker of clinical disease severity and mortality among antiretroviral therapy-naive Human Immunodeficiency Virus-infected adults in South Africa. Open Forum Infectious Diseases. 2017;4(3):ofx167. [DOI: 10.1093/ofid/ofx167].

LaCourse 2018a
LaCourse SM, Cranmer LM, Njuguna IN, Gatimu J, Stern J, Maleche-Obimbo E, et al. Urine tuberculosis lipoarabinomannan predicts mortality in hospitalized Human Immunodeficiency Virus-infected children. Clinical Infectious Diseases 2018;66(11):1798-1801. [DOI: 10.1093/cid/ciy011].

Lawn 2012b
Lawn SD, Kerkhoff AD, Vogt M, Wood R. Clinical significance of lipoarabinomannan detection in urine using a low-cost point-of-care diagnostic assay for HIV-associated tuberculosis. AIDS 2012;26(13):1635-43. [CRSREF: 2800857]

Lawn 2013
Lawn SD, Kerkhoff AD, Vogt M, Wood R. HIV-associated tuberculosis: relationship between disease severity and the sensitivity of new sputum-based and urine-based diagnostic assays. BMC Medicine 2013;11:231. [CRSREF: 2800861]

Manabe 2014

Manabe YC, Nonyane BA, Nakiyingi L, Mbabazi O, Lubega G, Shah M, et al. Point-of-care lateral flow assays for tuberculosis and cryptococcal antigenuria predict death in HIV infected adults in Uganda. PLoS One 2014;9(7):e101459. [CRSREF: 2800865]

Peter 2013
Peter JG, Theron G, Dheda K. Can point-of-care urine LAM strip testing for tuberculosis add value to clinical decision making in hospitalised HIV-infected persons? PLoS One 2013;8(2):e54875. [CRSREF: 2800881]

Additional references

Alere 2017
Alere Inc Waltham MA USA. Alere. Alere DetermineTM TB LAM Ag Product Information. 2017; IN02740004 Rev. 07 2017/08. https://www.alere.com/en/home/product-details/determine-tb-lam.html (accessed 17 Nov 2018)

Andrews 2014
Andrews B, Muchemwa L, Lakhi S, Bwalya M, Mabula C, Chipii G, Heimburger DC, Bernard GR. Validation of a clinical prediction score and -performance of urine lipoarabinomannan test for detecting tuberculosis bacteremia in HIV-positive patients with severe sepsis. American Thoracic Society 2014:A2559.

Balcha 2014
Balcha TT, Winqvist N, Sturegård E, Skogmar S, Reepalu A, Jemal ZH, et al. Detection of lipoarabinomannan in urine for identification of active tuberculosis among HIV-positive adults in Ethiopian health centres. Tropical Medicine & International Health 2014;19(6):734–42.

Balshem 2011
Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology. 2011;64(4):401-6.

Brennan 2003
Brennan P J. Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. Tuberculosis (Edinb) 2003;83(1-3):91-7.

Briken 2004
Briken V, Porcelli S A, Besra G S, Kremer L. Mycobacterial lipoarabinomannan and related lipoglycans: from biogenesis to modulation of the immune response. Molecular Microbiology 2004;53(2):391-403.

Broger 2018
Broger T, Sossen B, du Toit E, Kerkhoff AD, Schutz C, Reipold EI, et al. Novel high sensitivity tuberculosis point-of-care test for people living with HIV. Preprints with The Lancet (submitted not published), Available at SSRN: https://ssrn.com/abstract=3254479 2018.

Chu 2009
Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. Journal of the American Statistical Association 2009;104(486):512-23.

Covidence 2017
Covidence systematic review software. Available at www.covidence.org [Computer program]. Melbourne: Veritas Health Innovation, 2017.

Drain 2014a

Drain PK, Losina E, Coleman SM, Giddy J, Ross D, Katz JN, et al. Diagnostic accuracy of a point-of-care urine test for tuberculosis screening among newly-diagnosed HIV-infected adults: a prospective, clinic-based study. BMC Infectious Diseases 2014;**14**:110.

Drain 2014c
Drain P, Grobler A, Gounder L, Sahid F, Wilson D, Bassett I, et al. Rapid urine lipoarabinomannan testing after two months of tuberculosis treatment independently predicts mortality in a resource-limited setting. 44th World Conference on Lung Health of the International Union Against Tuberculosis and Lung Disease. Paris, 2013;17(Suppl 2):S264.

Ford 2016
Ford N, Matteelli A, Shubber Z, Hermans S, Meintjes G, Grinsztejn B, et al. TB as a cause of hospitalization and in-hospital mortality among people living with HIV worldwide: a systematic review and meta-analysis. Journal of the International AIDS Society 2016;19(1):20714.

Gupta 2015
Gupta RK, Lucas SB, Fielding KL, Lawn SD. Prevalence of tuberculosis in post-mortem studies of HIV-infected adults and children in resource-limited settings: a systematic review and meta-analysis. AIDS 2015;29(15):1987-2002.

Gupta-Wright 2016
Gupta-Wright A, Peters JA, Flach C, Lawn SD. Detection of lipoarabinomannan (LAM) in urine is an independent predictor of mortality risk in patients receiving treatment for HIV-associated tuberculosis in sub-Saharan Africa: a systematic review and meta-analysis. BMC Med 2016;14:53.

Gupta-Wright 2018
Gupta-Wright A, Kerkhoff AD, Meintjes G, Corbett EL. Urinary lipoarabinomannan detection and disseminated nontuberculous mycobacterial disease. Clinical Infectious Diseases 2018;66(1):158.

Gupta-Wright 2018a
Gupta-Wright A, Corbett EL, Oosterhout JJ, Wilson D, Grint D, et al. Rapid urine-based screening for tuberculosis in HIV-positive patients admitted to hospital in Africa (STAMP): a pragmatic, multicentre, parallel-group, double-blind, randomised controlled trial. Lancet 2018; S0140-6736(18)31267-4.

Harris 2009
Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics
2009;42(2):377-81.

Kohli 2018
Kohli M, Schiller I, Dendukuri N, Dheda K, Denkinger CM, Schumacher SG, et al. Xpert((R)) MTB/RIF assay for extrapulmonary tuberculosis and rifampicin resistance. Cochrane Database of Systematic Reviews 2018;8:CD012768.

Lawn 2012
Lawn S D. Point-of-care detection of lipoarabinomannan (LAM) in urine for diagnosis of HIV-associated tuberculosis: a state of the art review. BMC Infectious Diseases 2012;12(1):103.

Lawn 2012a
Lawn SD, Kerkhoff AD, Vogt M, Wood R. Diagnostic accuracy of a low-cost, urine antigen, point-of-care screening assay for HIV-associated pulmonary tuberculosis before antiretroviral therapy: a descriptive study. Lancet Infectious Diseases 2012;**12**(3):201–9.

Lawn 2013a
Lawn SD, Dheda K, Kerkhoff AD, Peter JG, Dorman S, Boehme CC, et al. Determine TB-LAM lateral flow urine antigen assay for HIV-associated tuberculosis: recommendations on the design and reporting of clinical studies. BMC Infectious Diseases 2013;13:407.

Lawn 2014a
Lawn SD, Kerkhoff A, Burton R, Schutz C, van Wyk G, Vogt M, et al. Massive diagnostic yield of HIV-associated tuberculosis using rapid urine assays in South Africa. Conference on Retroviruses and Opportunistic Infections (CROI). Boston (MA), 2014.

Lawn 2015
Lawn SD, Kerkhoff AD, Nicol PM, Meintjes G. Underestimation of the true specificity of the urine lipoarabinomannan (LAM) point-of-care diagnostic assay for HIV-associated tuberculosis. Journal of Acquired Immune Deficiency Syndromes 2015;69(4):e144-6.

Lawn 2016
Lawn SD, Gupta-Wright A. Detection of lipoarabinomannan (LAM) in urine is indicative of disseminated TB with renal involvement in patients living with HIV and advanced immunodeficiency: evidence and implications. Transactions of the Royal Society of Tropical Medicine and Hygiene 2016;110(3):180-5.

Lunn 2009
Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. Statistics in Medicine. 2009;28(25):3049-67.

Macaskill 2010
Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from http://srdta.cochrane.org/.

Minion 2011
Minion J, Leung E, Talbot E, Dheda K, Pai M, Menzies D. Diagnosing tuberculosis with urine lipoarabinomannan: systematic review and meta-analysis. European Respiratory Journal 2011;38(6):1398-405.

Nel 2017
Nel J S, Lippincott C K, Berhanu R, Spencer D C, Sanne I M, Ive P. Does disseminated nontuberculous mycobacterial disease cause false-positive determine TB-LAM lateral flow assay results? A retrospective review. Clinical Infectious Diseases 2017;65(7):1226-8.

Peter 2010
Peter J, Green C, Hoelscher M, Mwaba P, Zumla A, Dheda K. Urine for the diagnosis of tuberculosis: current approaches, clinical applicability, and new developments. Current Opinion of Pulmonary Medicine 2010;16(3):262-70.

Qvist 2014
Qvist T, Johansen I S, Pressler T, Hoiby N, Andersen AB, Katzenstein T L, et al. Urine lipoarabinomannan point-of-care testing in patients affected by pulmonary nontuberculous mycobacteria - experiences from the Danish Cystic Fibrosis cohort study. BMC Infectious Diseases 2014;14:655.

R Statistical Computing 2015

R: A language and environment for statistical computing [Computer program]. R Development Core Team. Version 3.2.1. Vienna: R Foundation for Statistical Computing,2015.

Reitsma 2005
Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005;58(10):982-90.

Review Manager
Review Manager (RevMan) [Computer program]. Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

Sabur 2017
Sabur NF, Esmail Aliasgar, Brar MS, Dheda K. Diagnosing tuberculosis in hospitalized HIV-infected individuals who cannot produce sputum: is urine lipoarabinomannan testing the answer? BMC Infectious Diseases 2017;17:803.

Schünemann 2008
Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336(7653):1106-11.

Schünemann 2016
Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Working Group. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. Journal of Clinical Epidemiology. 2016 Aug;76:89-98.

OPEN, Odense Patient data Explorative Network, Odense University Hospital, Odense, Denmark https://www.sdu.dk/ki/open (Accessed 10 Jun 2018).

Shah 2010
Shah M, Martinson NA, Chaisson RE, Martin DJ, Variava E, Dorman SE. Quantitative analysis of a urine-based assay for detection of lipoarabinomannan in patients with tuberculosis. Journal of Clinical Microbiology. 2010;48(8):2972-4.

Spiegelhalter 2004
Spiegelhalter DJ, Abrams KR., Myles JP. Bayesian approaches to clinical trials and health-care evaluation [Bayesian approaches to clinical trials and health-care evaluation]. Wiley 2004.

StataCorp 2017
Stata Statistical Software [Computer program]. StataCorp. Version 13. Texas: StataCorp LP, 2017.

TB CARE I 2014
TB CARE I. International standards for tuberculosis care, Edition 3. TB Care 1, The Hague. http://www.who.int/tb/publications/ISTC_3rdEd.pdf?ua=1 (accessed 16 Nov 2014);1-87.

Whiting 2011
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. Annals of Internal Medicine 2011;155(8):529–36.

WHO End TB 2014
World Health Organization. The End TB strategy. Global strategy and targets for tuberculosis prevention, care and control after 2015. 2014.

WHO Global Report 2018
World Health Organization. Global Tuberculosis Report. 2018.

WHO Lipoarabinomannan Policy Guidance 2015
World Health Organization. The use of lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis and screening of active tuberculosis in people living with HIV. Policy guidance. WHO/HTM/TB/2015.25. Geneva: World Health Organization, 2015.

WHO TTP 2014
World Health Organization. High-priority target product profiles for new tuberculosis diagnostics, report of a consensus meeting. 2014;98.

World Bank 2018/2019
World Bank. World Bank Country Classifications by Income. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed 15 May 2018).

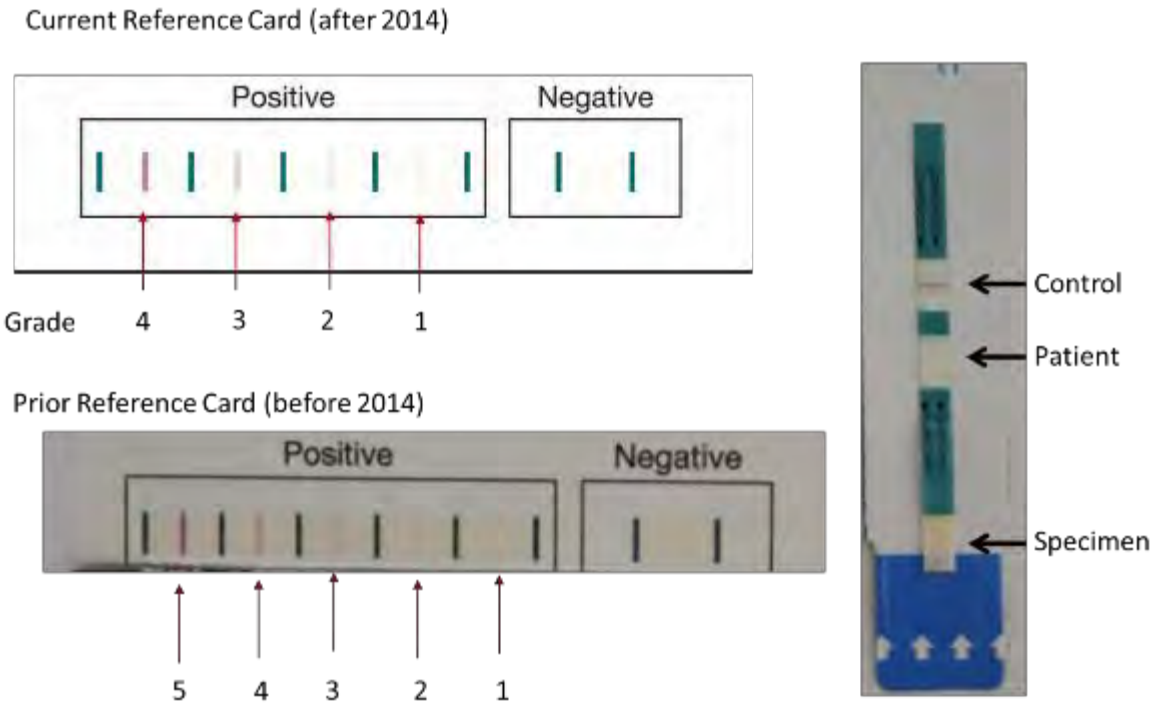
Other published versions of this review

Shah 2014
Shah M, Hanrahan C, Wang ZY, Steingart KR, Lawn SD, Denkinger C, et al. Urine lateral flow lipoarabinomannan assay for diagnosing active tuberculosis in adults living with HIV [Protocol]. Cochrane Database of Systematic Reviews 2014, Issue 12. Art. No.: CD011420. DOI: 10.1002/14651858.CD011420.

Shah 2016
Shah M, Hanrahan C, Wang ZY, Dendukuri N, Lawn SD, Denkinger CM, Steingart KR. Lateral flow urine lipoarabinomannan assay for detecting active tuberculosis in HIV-positive adults. Cochrane Database of Systematic Reviews 2016, Issue 5. Art. No.: CD011420. Art. No.: CD011420. DOI: 10.1002/14651858.CD011420.pub2.

Appendices

Appendix 1. Reference card grading of Alere Determine™ TB LAM



Appendix 2. PICO questions

What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in all HIV-positive adults and children with signs and symptoms of TB?

in inpatient settings (adults, adolescents and older children)
in outpatient settings (adults, adolescents and older children)
all settings (adults, adolescents and older children)
in inpatient settings (children ≤ 5 years)
in outpatient settings (children ≤ 5 years)
all settings (children ≤ 5 years)

What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in all HIV-positive adults and children irrespective of signs and symptoms of TB?

in inpatient settings (adults, adolescents and older children)
in outpatient settings (adults, adolescents and older children)
all settings (adults, adolescents and older children)
in inpatient settings (children ≤ 5 years)
in outpatient settings (children ≤ 5 years)
all settings (children ≤ 5 years)

What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in adults with advanced HIV disease irrespective of signs and symptoms of TB?

in inpatient setting CD4 ≤ 200
in outpatient setting CD4 ≤ 200
in all settings CD4 ≤ 200
in inpatient setting CD4 ≤ 100
in outpatient setting CD4 ≤ 100
in all settings CD4 ≤ 100

Can the use of LF-LAM in HIV-positive adults reduce mortality associated with advanced HIV disease?

in all settings
in inpatient settings
in outpatient settings
in individuals with CD4 ≤ 200
in inpatient setting CD4 ≤ 200
in outpatient setting CD4 ≤ 200
in individuals with CD4 ≤ 100
in inpatient setting CD4 ≤ 100
in outpatient setting CD4 ≤ 100

Other questions: What is the cost and cost-effectiveness of LF-LAM implementation for TB diagnosis, based on review of the published literature?

Appendix 3. Detailed search strategies

MEDLINE (Pubmed) search history

Search	
#9	Search (#3) AND (#7) AND #8)
#8	Search test OR assay OR antigen OR Ag OR lateral flow assay*OR urine antigen OR point of care Field: Title/Abstract
#7	Search (#4) OR #5) OR #6
#6	Search LAM; Field: Title/Abstract
#5	Search "lipoarabinomannan" [Supplementary Concept]
#4	Search lipoarabinomannan ; Field: Title/Abstract
#3	Search (#1) OR #2)
#2	Search tuberculosis Or TB Field: Title/Abstract
#1	Search ("Tuberculosis"[Mesh]) OR "Mycobacterium tuberculosis"[Mesh]

Database: EMBASE 1947-Present, updated daily

Search Strategy:

---

1 tuberculosis.mp. or tuberculosis/ or Mycobacterium tuberculosis/ (115438)
2 limit 1 to yr="2014 -Current" (8833)
3 lipoarabinomannan.mp. or lipoarabinomannan/ (775)
4 LAM.mp. (4928)
5 limit 4 to yr="2014 -Current" (500)
6 3 or 5 (1252)
7 2 and 6 (79)
8 (test or assay or antigen or Ag or lateral flow assay* or urine antigen or point of care).mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] (2733052)
9 limit 8 to yr="2014 -Current" (223771)
10 7 and 9 (46)

Cochrane Central Register of Controlled Trials: Issue 4 of 12, April 2018

ID Search
#1 tuberculosis:ti,ab,kw (Word variations have been searched)
#2 TB:ti, ab, kw
#3 MeSH descriptor: [Mycobacterium tuberculosis] explode all trees
#4 MeSH descriptor: [Tuberculosis] explode all trees
#5 #1 or #2 or #3 or #4
#6 LAM:ti,ab,kw
#7 lipoarabinomannan:ti,ab,kw
#8 #6 or #7
#9 #5 and #8

Web of Science Core Collection - Indexes: SCI-EXPANDED, CPCI-S, Biosis previews

TOPIC: (tuberculosis OR TB OR mycobacterium) *AND* **TOPIC**: (lipoarabinomannan OR LAM) *AND* **TOPIC**: (test OR assay OR antigen OR Ag OR lateral flow assay* OR urine antigen OR point of care)

SCOPUS

(TITLE-ABS-KEY (tuberculosis OR TB) AND TITLE-ABS-KEY (lipoarabinomannan OR LAM) AND (test OR diagnos* OR urine OR assay)

CIDG Specialized Register, LILACS, Proquest dissertations, Current Controlled trials, WHO trials register:

Tuberculosis AND (lipoarabinomannan OR LAM)

Appendix 4. Data collection form, diagnostic accuracy

AlereLAM - Lateral flow urine lipoarabinomannan assay for diagnosing active tuberculosis in people living with HIV Data form		
	I. STUDY IDENTIFICATION	
1	First author	
2	Corresponding author and email	
3	Title of study	
4	Year of publication	
5	Year of study start	
6	Language if other than English	
	II. STUDY DETAILS	
7	Population	1.Adults (15 years of age) 2.Children and adolescents 3.Both adults, children and adolescents 4.Other If other, describe
8	In which country or countries was the study conducted?	List all countries:
9	Country World Bank Classification (income)	1.Low income 2.Lower-middle income 3.Upper-middle-income 4.High income 7.Other combination 9.Unknown/Not reported If other, describe:
10	Country WHO classification for high TB burden country (WHO 2015)	1.Yes, part of the High TB/HIV burden list 2.No, not part of the High TB/HIV burden list
11	Study design	1.Randomized controlled trial 2.Cross-sectional 3.Cohort 7.Other, specify 9.Could not tell If other, describe:
12	Was a case-control design avoided?	1.Yes 2.No 9.Unclear
	III. PATIENT SELECTION	
13	What was the manner of participant selection into the study?	1.Consecutive 2.Random 3.Convenience 7.Other, specify 9.Unknown/Not Reported/Unclear If other, describe:

14	Direction of study data collection	1.Prospective 2.Retrospective 9.Unknown/Not reported
15	Please select the statement that best describes the selection of participants into the study.	1.HIV-positive participants with signs or symptoms suggestive of active TB were tested using AlereLAM. Please provide study definition of 'signs and symptoms': 2.A predetermined target population of HIV-positive individuals, irrespective of signs and symptoms of TB, were tested using AlereLAM. Please specify target population: 3.Both 1 and 2 4.Neither 1 nor 2. This is what was done:
16	Sample size	1._____9.Unknown/Not reported
17	Did the study avoid inappropriate exclusions?	1.Yes 2.No 9.Unknown/Not reported/Unclear
17a	Could the selection of patients have introduced bias?	1.High risk 2.Low risk 9.Unclear risk
18	NOTES ON PATIENT SELECTION	
	IV. PATIENT CHARACTERISTICS AND SETTING	
19	Presenting signs and symptoms	List
20	Age (years)	_____ If age is reported in median indicate IQR If age is reported in mean indicate SD
21	Age of all study participants, Range	Upper Lower
22	HIV infection (%)	
23	Participants included of female sex (%)	
24	CD4	_____ If CD4 is reported in median indicate IQR If CD4 is reported in mean indicate SD
25	Number (percent) of TB cases in the study (%):	
26	What was the target condition?	1.Pulmonary TB 2.Extra pulmonary TB 3.Mycobacteraemia 4.Both 1 and 2 5.Any of 1,2,3 7.Other, specify
27	Did the study include patients with prior TB history?	1.Yes 2.No 9.Unknown/Not reported If yes, what is the % _____ Specify the numerator/denominator_____/_____

28	What was the clinical setting of the study?	1.Outpatient 2.Inpatient 3.Both out-patient and in-patient 7.Other, describe: 9.Unknown/Not reported
29	How would you describe the health facility where the study took place?	1.Primary care clinic, stand-alone 2.Primary care clinic, connected to a referral hospital 3.Referral hospital 7.Other, describe: 9.Unknown/Not reported
30	Are there concerns that the included patients and setting do not match the review question?	1.High concern 2.Low concern 9.Unclear concern
31	NOTES ON CHARACTERISTICS	
	V. INDEX TEST	
32	Was a AlereLAM threshold used to define positivity that was pre-specified in the primary analysis?	1.Yes, Grade 1/5 2.Yes, Grade 2/5 3.Yes, Grade ¼ 4.Yes, Grade 2/4 5.No 7.Other, specify:
33	What AlereLAM threshold was used to define positivity for data extraction?	1.Grade 2/5 2.Grade 1/4 7. Other, specify
34	Are their concerns about index test conduct or interpretation differing from review question?	1.High concern 2.Low concern 9.Unclear concern
35	Was AlereLAM performed on fresh or stored urine?	1.Fresh 2.Stored, specify type of storage (e.g. frozen) 3.Both fresh and stored 9.Unknown/Not reported
36	Was AlereLAM result interpreted without knowledge of the result of the reference standard result?	1.Yes 2.No 9.Unknown/Not reported/Unclear
37	Were there any AlereLAM results that were invalid (no bar in control window)?	1.Yes a.Specify number of invalid tests: _____ b.Were invalid tests repeated (yes/no): _____ 2.No 9. Unknown/Not reported
38	Could the conduct or interpretation of the index test have introduced bias?	1.High risk 2.Low risk 9.Unclear risk
39	NOTES ON INDEX TEST	
	VI. REFERENCE STANDARD	
40	For the diagnosis of pulmonary TB, what reference standard was used to identify TB and not TB?	1.Sputum: solid culture 2.Sputum: liquid culture 3.Sputum: both solid and liquid culture

		4.Nucleic acid amplification test, specify 5.Any of culture or nucleic amplification test, specify 7.Other, specify
41	Was sputum induction performed for individuals unable to produce expectorated sputum?	1.Yes Specify N/% requiring sputum induction _____ 2.No
42	Were patients without sputum specimens (for example, no expectorated, no induced sputum) included in this study?	1.Yes Specify N/% included without sputum _____ 2.No Specify N/% excluded due to lack of sputum_____
43	Were non-pulmonary specimens evaluated to allow diagnosis of extrapulmonary TB?	1.All participants received testing of non-pulmonary specimens, please specify sites/fluids: 2.Some participants received testing of non-pulmonary specimens, please specify which patients were tested, and sites/fluids: 3.Extrapulmonary TB was not evaluated 7.Other, please specify:
44	For the diagnosis of extrapulmonary TB, what tests were used to identify TB and not TB (circle all that apply)?	1.Solid culture 2.Liquid culture 3.Both solid and liquid culture 4.Nucleic acid amplification test, specify 7.Other, specify: _____ 8.Not applicable, extrapulmonary TB was not evaluated
45	Did the study speciate mycobacteria isolated in culture?	1.Yes 2.No 9.Unknown/Not reported
46	Was the reference standard likely to correctly classify the target condition	1.Yes 2.No 9. Unclear
47	Was the reference standard result interpreted without knowledge of the result of AlereLAM?	1.Yes 2.No 9.Unclear
48	How many sputum specimens were obtained in order to detect pulmonary TB?	1. Single 2. Multiple 8.Not applicable
49	How many specimens from fluid (sites) other than sputum were obtained to detect extrapulmonary TB?	1. Single 2. Multiple 8.Not applicable
50	Could the reference standard, its conduct, or its interpretation have introduced bias?	1.High risk 2.Low risk 9.Unclear risk
51	Are there concerns that the target condition as defined by the reference standard does not match the question?	1.High concern 2.Low concern 9.Unclear concern

52	NOTES ON REFERENCE STANDARD	
	VII. FLOW AND TIMING	
53	Was there appropriate interval between index test and reference standard	1.Yes, specimens collected at the same time. 2.No, specimens collected greater than 7 days apart 9.Unclear
54	Did all patients receive a reference standard?	1.Yes 2.No 9.Unclear
55	Did all patients receive the same reference standard?	1.Yes 2.No (answer no if clinicians chose sample types, or other differences in reference standards between patients) 9.Unclear
56	Were all participants included in the analysis?	1.Yes 2.No 9.Unclear
57	Could the patient flow have introduced bias?	1.High risk 2.Low risk 9.Unclear risk
58	NOTES ON FLOW AND TIMING	

Abbreviations: IQR, interquartile range; SD, standard deviation.

VIII. TABLES: TB detection against a microbiological reference standard
TB is defined as positive culture or NAAT from sputum or any other body fluid or site.
Not TB is defined as negative cultures or NAATs from sputum or any other body fluid or site.

(Table example to extract TP, FP, FN, TN values)

LAM result		**TB**	**Not TB**	**Total**
	Positive			
	Negative			
	Total			

Provide additional tables for each of the applicable PICO questions 1-4 (Appendix 2. PICO questions) and the following additional questions:

5. What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in all HIV-positive adults with advanced HIV disease and signs and symptoms of TB?

a. in inpatient setting CD4≤ 200
b. in outpatient setting CD4≤ 200
c. in all settings CD4≤ 200
d. in all settings CD4 > 200
e. in inpatient setting CD4≤ 100
f. in outpatient setting CD4≤ 100
g. in all settings CD4≤ 100
h. in all settings CD4 > 100
i. in inpatients settings CD4 101-200
j. in outpatient settings CD4 101-200
k. in all settings CD4 101-199

6. What is the diagnostic accuracy of LF-LAM for the diagnosis of TB in adults with advanced HIV disease irrespective of signs and symptoms of TB?

a. in all settings CD4 > 200
b. in all settings CD4 > 100
c. in inpatients settings CD4 101-200
d. in outpatient settings CD4 101-200
e. in all settings CD4 101-199

Appendix 5. Data collection form, impact data

AlereLAM - Lateral flow urine lipoarabinomannan assay for diagnosing active tuberculosis in people living with HIV Data form for impact data extraction		
	I. STUDY IDENTIFICATION	
1	First author	
2	Journal	
3	Year of publication	
	II. STUDY DETAILS	
4	Population	Adults or children HIV status Other details re: study inclusion criteria
5	In which country or countries was the study conducted?	List all countries:
6	Study design	Randomized controlled trial Cohort Cross- sectional Other
	III. PATIENT SELECTION	
7	Direction of study data collection	Prospective Retrospective Unknown/Not reported
8	Please select the statement that best describes the selection of participants into the study.	HIV-positive participants with signs or symptoms suggestive of active TB were tested using AlereLAM. Please provide study definition of 'signs and symptoms': A predetermined target population of HIV-positive individuals, irrespective of signs and symptoms of TB, were tested using AlereLAM. Please specify target population: Both 1 and 2 Neither 1 nor 2. This is what was done:
9	Number enrolled	
10	Number in analysis	
	IV. PATIENT CHARACTERISTICS AND SETTING	
11	Age (years)	 If age is reported in median indicate IQR If age is reported in mean indicate SD
12	What was the clinical setting of the study?	Outpatient Inpatient Both out-patient and in-patient

		Other, describe: Unknown/Not reported
	V. INDEX TEST (LAM)	
13	What AlereLAM threshold was used to define positivity for data extraction?	Grade 2/5 Grade 1/4 Other, specify
14	Was old or new AlereLAM card used	Old (5 grades) New (4 grades)
	VI. MORTALITY ASSESSMENT	
15	How was mortality assessed?	Describe
16	Type of mortality?	All-cause TB-related
17	When was mortality assessed?	
18	Mortality (ARR)	Describe results
19	Mortality (HR, aHR, MHR, or Kaplan Meier)	Describe results
20	Mortality (OR or aOR)	Describe results
21	What was the comparator?	Describe
22	Mortality in intervention	Percentage and number
23	Mortality in control	Percentage and number
24	Mortality in LAM positive	Percentage and number
25	Mortality in LAM negative	Percentage and number
26	Mortality in LAM positive confirmed TB cases	Percentage and number
27	Mortality in LAM negative confirmed TB cases	Percentage and number
28	Mortality in LAM positive patients with an inconclusive evaluation for TB	Percentage and number
29	Mortality in LAM negative patients with an inconclusive evaluation for TB	Percentage and number
30	Mortality in LAM positive non-TB cases	Percentage and number
31	Mortality in LAM negative non-TB cases	Percentage and number
32	Was analysis stratified by CD4 count or percentage?	Describe/list results if stratified by CD4 count or percentage
33	Time to diagnosis	Number of days
34	Time to treatment	Number of days
35	Other outcomes assessed in the study	Describe
36	Comments	Describe

Appendix 6. QUADAS-2

Domain 1: patient selection

Risk of bias: could the selection of patients have introduced bias?

Signalling question 1: Was a consecutive or random sample of patients enrolled?
We answered 'yes' if the study enrolled a consecutive or random sample of eligible participants; 'no' if the study selected participants by convenience; and 'unclear' if the study did not report the manner of participant selection or we could not tell.

Signalling question 2: Was a case-control design avoided?
We answered 'yes' to all included studies given that we are excluding case-control study designs.

Signalling question 3: Did the study avoid inappropriate exclusion?
We answered 'yes' to studies which included all HIV-positive participants and participants who were unable to produce sputum (expectorated or induced). We answered 'no' if studies excluded participants who could not produce sputum (i.e. there were no attempts at sputum induction or patients could not produce sputum despite sputum induction and were excluded). We also answered 'no' if studies excluded patients presumed to have extrapulmonary TB. We scored 'unclear' if we could not tell

Applicability: Are there concerns that the included patients and setting do not match the review question?

We were interested in how AlereLAM performs in patients whose urine specimens were evaluated as they would be in routine practice. We expected to judge ‘low concern’ for most studies since we planned to determine test accuracy both for patients with signs and symptoms of TB and patients investigated for TB irrespective of signs and symptoms for TB.

For AlereLAM used as a TB diagnostic test among patients with signs and symptoms of TB, we judged 'high concern' if the study participants did not resemble people with presumed HIV/TB; 'low concern' if the study population did resemble a population with presumed HIV/TB, and 'unclear concern', if we could not tell.

For AlereLAM used as a TB diagnostic test among patients that were investigated for TB irrespective of signs and symptoms of TB, we judged 'low concern' for studies in which the AlereLAM was performed uniformly within the predetermined study target populations of HIV-infected individuals, 'high concern' if AlereLAM was not performed uniformly within the predetermined study target populations of HIV-infected individuals, and 'unclear concern' if we could not tell. We judged 'high concern' if the study participants did not resemble people with presumed HIV/TB coinfection.

Domain 2: index test

Risk of bias: could the conduct or interpretation of the index test have introduced bias?

Signalling question 1: were the index test results interpreted without knowledge of the results of the reference standard?
We answered 'yes' if the study interpreted the result of AlereLAM blinded to the result of the reference standard; we answered 'no' if the study did not interpret the result of AlereLAM blinded to the result of the reference standard. We answered 'yes' for studies in which AlereLAM was performed on fresh specimens, since reference standard results would be unavailable at the time of test interpretation. We answered 'unclear' if stored specimens were tested and we could not tell if the index test results were interpreted without knowledge of the reference standard results.

Signalling question 2: if an AlereLAM threshold was used to define positivity, was it prespecified?

We answered 'yes' if the threshold was prespecified in the study or by the authors, 'no' if the threshold was not prespecified, and 'unclear' if we could not determine if the threshold was prespecified or not.

Applicability: are there concerns that the index test, its conduct, or its interpretation differ from the review question?

If index test methods vary from those specified in the review question, concerns about applicability may exist. We judged 'high concern' if the test procedure was inconsistent with the manufacturer recommendations, 'low concern' if the test procedure was consistent with the manufacturer recommendations, and 'unclear concern' if we could not tell. In cases where the primary study defined grade 1 of 5 as the positivity threshold, but where we were able to extract data at the manufacturer's currently recommended positivity threshold, we judged 'low concern' for applicability.

Domain 3: reference standard

Risk of bias: could the reference standard, its conduct, or its interpretation have introduced bias?

Signalling question 1: is the reference standard likely to correctly classify the target condition?

HIV-infected TB patients may have pulmonary TB, extrapulmonary TB, or both pulmonary and extrapulmonary TB. A microbiological reference standard, primarily culture, is considered the gold standard for TB. Due to the difficulties in diagnosing HIV-associated TB, it is recommended that multiple cultures from sputum and other specimens be evaluated.

We answered 'yes' when appropriate specimens were obtained for the diagnosis of HIV-associated TB. For presumed pulmonary TB, sputum specimens should be obtained for culture, NAAT, or both culture and NAAT. If the patient cannot produce sputum, induced sputum should be performed. For presumed extrapulmonary TB, specimens should be consistent with Standard 4 of the International Standards for TB Care which states: "For all patients, including children, suspected of having extrapulmonary tuberculosis, appropriate specimens from the suspected sites of involvement should be obtained for microbiological and histological examination" (TB CARE I 2014). We answered yes if multiple specimens were collected from different sites for extrapulmonary TB. An Xpert® MTB/RIF test is recommended as the preferred initial microbiological test for suspected TB meningitis because of the need for a rapid diagnosis". We also answered 'yes' if studies followed a standardized approach of collecting appropriate specimens from "suspected sites of involvement", for example, blood or lymph nodes on all patients.
We answered 'no' when the reference standard was restricted to sputum specimens or the reference standard was restricted to extrapulmonary specimens (for example, urine, blood, etc.). We also answered 'no' if a consistent approach was not followed for all patients (for example, some but not all patients with presumed TB lymphadenitis receive lymph node tissue sampling). We answered 'unclear' if we could not tell.

Signalling question 2: were the reference standard results interpreted without knowledge of the results of the index test?

We answered 'yes' if the study interpreted the result of the reference standard blinded to the result of AlereLAM, or if the reference standard result was reported on an automated instrument; 'no' if the study did not interpret the result of the reference standard blinded to the result of AlereLAM, and 'unclear' if we could not tell.

Applicability: are there concerns that the target condition as defined by the reference standard does not match the question?

In general, we thought there was low concern for almost included studies based on the current definitions of the reference standard. We judged 'high concern' if included studies did not speciate mycobacteria isolated in culture, 'low concern' if speciation was performed, and 'unclear' if we could not tell. We also judged high concern if there was no protocol to ensure a minimum standard of testing with a reference standard.

Domain 4: Flow and timing

Risk of bias: could the patient flow have introduced bias?

Signalling question 1: was there an appropriate interval between the index test and reference standard?

We expected urine specimens for AlereLAM and the reference standards to be obtained at the same time and answered 'yes' for all studies that meet this criterion, or if index and reference standard tests were performed on specimens collected no greater than seven days apart. We chose seven days as a time period during which either treatment of TB or natural progression of TB without treatment could impact test results. We answered 'no' if specimens were collected for index and reference standard tests greater than seven days apart, and 'unclear' if we could not tell.

Signalling question 2: did all patients receive the same reference standard?

We answered 'yes' if all participants in the study received the reference standard to confirm TB; 'no' if not all patients received the reference standard to confirm TB, and 'unclear' if we could not tell.

Signalling question 3: were all patients included in the analysis?

We determined the answer to this question by comparing the number of participants enrolled in the study with the number of participants included in the two-by-two tables. We answered 'yes' if all participants enrolled in the study were tested with results presented and accounted for. We answered 'no' if participants meeting enrolment criteria were not tested or results were not presented, and 'unclear' if we could not tell.

Judgements for 'Risk of bias' assessments

If we answered all signalling questions for a domain "yes", then we judged risk of bias as "low".
If we answered all or most signalling questions for a domain "no", then we judged risk of bias as "high".
If we answered only one signalling question for a domain "no", we discussed further the "risk of bias" judgement.
If we answered all or most signalling questions for a domain "unclear", then we judged risk of bias as "unclear".
If we answered only one signalling question for a domain "unclear", we discussed further the "risk of bias" judgement for the domain.

Appendix 7. Statistical approach

We list here the OpenBUGS program used to fit the bivariate meta-analysis models for estimating the accuracy of the index test. In the subsections below, we first describe the likelihood and prior distribution for the model followed by the OpenBUGS program.

As is usual with Bayesian models, initial values must be provided for all unknown parameters. We selected three independent sets of initial values for the parameters using the in-built ModelGenInits() function within OpenBUGS. The Gelman-Rubin statistic within the OpenBUGS program was used to assess convergence. We did not observe any convergence problems for the analyses presented. We treated the first 10,000 iterations as burn-in iterations and dropped them. We obtained summary statistics based on a total of 150,000 iterations resulting from the three separate chains.

A. Estimation of index test accuracy

Notation: in the i-th study the cells in the cross-tabulation between the index and reference tests are denoted by TPi, FPi, TNi, FNi. The sensitivity in i-th study is denoted by sei and the specificity by spi.

We denote the Binomial probability distribution with sample size N and probability p as Binomial(p,N), the Bivariate Normal probability distribution with mean vector μ and variance-covariance matrix TAU as BVN(mu, TAU), the univariate Normal distribution with mean m and variance tau2 by N(m, tau2) and the Uniform probability distribution between a and b by Uniform(a,b). Note that logit refers to log odds.

Likelihood:

Within studies:
TPi ~ Binomial(TPRi, TPi + FNi), and
FPi ~ Binomial(FPRi, TNi + FPi)

Between studies:
The bivariate vector (logit(TRPi), logit(FPRi)) ~ BVN(mu = (mu1, mu2), TAU) where
TAU is a 2 X 2 matrix with entries

TAU[1,1] = variance of logit(TPRi) = tau12,

TAU[2,2] = variance of logit(FPRi) = tau22 and

TAU[1,2] = TAU[2,1] = covariance between logit(TPRi) and logit(FPRi) = rho × tau1 × tau2

and rho is the correlation between logit(TPRi) and logit(FPRi) across studies.

The pooled sensitivity is given by 1/(1+exp(-mu1)), and the pooled specificity is given by 1/(1+exp(-mu2)).

Prior distributions:

mu1 and mu2 ~ N(m=0, tau2=4),

rho ~ Uniform(-1, 1)

(1/ tau12) and (1/ tau22) ~ Gamma(shape=2, rate=0.5)

A.1 OpenBUGS program for estimating a bivariate hierarchical meta-analysis model for sensitivity and specificity of the index test.

Observed data must be provided for L (the number of studies), and TP, FN, FP and TN in each study.

```
model {
for(i in 1:L) { ## L is the number of studies in the Meta-analysis
# Likelihoood
pos[i]<-TP[i]+FN[i]
neg[i]<-TN[i]+FP[i]
TP[i] ~ dbin(TPR[i],pos[i])
FP[i] ~ dbin(FPR[i],neg[i])logit(TPR[i]) <- l[i,1]
logit(FPR[i]) <- -l[i,2]se[i] <- TPR[i]
sp[i] <- 1-FPR[i]
l[i,1:2] ~ dmnorm(mu[1:2], T[1:2, 1:2])
}
# Prior Distributions
mu[1] ~ dnorm(0,0.25)
mu[2] ~ dnorm(0,0.25)
T[1:2,1:2]<-inverse(TAU[1:2,1:2])
# Between-study variance-covariance matrix
TAU[1,1] <- tau[1]*tau[1]
TAU[2,2] <- tau[2]*tau[2]
TAU[1,2] <- rho*tau[1]*tau[2]
TAU[2,1] <- rho*tau[1]*tau[2]
# prec is the between-study precision in the logit(sensitivity) and logit(specificity)
# rho is the correlation between logit(sensitivity) and logit(specificity) across studies
prec[1] ~ dgamma(2,0.5)
prec[2] ~ dgamma(2,0.5)
rho ~ dunif(-1,1)
tau[1]<-pow(prec[1],-0.5)
```

```
tau[2]<-pow(prec[2],-0.5)

# Pooled sensitivity and specificity

Pooled_S<-1/(1+exp(-mu[1]))

Pooled_C<-1/(1+exp(-mu[2]))

}
```

Appendix 8: Characteristics of Included Studies

Study	Participants (% symptomatic)	Setting	Median CD4 cell count per µL (IQR)	TB prevalence % (n/N)	Did the study avoid inappropriate exclusion	Specimens collected	High quality reference standard*	Unique Study Characteristics
HIV positive adults with signs and symptoms of TB								
Drain 2016	**Symptomatic:** Two of four TB related symptoms (cough, fever weight loss, night sweat) for > 2 weeks; smear microscopy negative x 2	Outpatient	168 (89-256)	63% (57/90)	No	Pulmonary samples	No	Adults (>18 years); HIV positive (93.2%); Targeting a relatively well outpatient population; Karnofsky performance score >50
Huerga 2017	**Symptomatic:** Cough > 2 weeks or any cough and one of weight loss, night sweats or fever; severely ill; CD4< 200 or BMI below 17	Outpatients (33%); Inpatients (67%)	109 (43-214)	57% (156/275)	No	Pulmonary samples; Urine Xpert only for patients without sputum available	No	Adults (>15 years); LAM guided treatment; Excluded many participants from analysis**
Juma 2017	**Symptomatic:** Suggestive of extrapulmonary TB, not specified	Inpatients	not stated	33% (29/67)	No	Extrapulmonary samples only, no sputum samples	No	Adults (>14 years), HIV- positive (68%); Excluded patients with concomitants active pulmonary TB
Nakiyingi 2014	**Symptomatic:** Any of cough, fever weight loss, night sweat	Outpatients (45%); Inpatients (55%)	152 (41-337)	37% (367/997)	No	Pulmonary samples; Blood culture for all	Yes	Adults (>18 years); multisite; large sample size.
Pandie 2016	**Symptomatic:** Presence of a pericardial effusion and suspected of pericardial TB	Inpatients	139 (81-249)	95% (36/38)	No	Extrapulmonary samples (pericardial effusion); pulmonary samples for some	No	Adults (>18 years); HIV-positive (74%); Excluded participants from analysis affecting specificity***.
Peter 2012	**Symptomatic**: Any of cough, fever weight loss, night sweat	Inpatients	90 (47-197)	48% (116/241)	Yes	Clinically relevant pulmonary samples; clinically relevant extrapulmonary samples. No study defined algorithm.	No	Adults (> 18 years). Multisite; TB diagnostic work-up was not standardised but up to clinical judgements
Peter 2015	**Symptomatic:** Any of cough, fever weight loss, night sweat	Outpatient	210 (103-375)	32% (181/569)	No	Pulmonary samples	No	Adults (> 18 years), Multisite; nested within a randomised, parallel-arm trial.
Peter 2016	**Symptomatic**: Any of cough, fever weight loss, night sweat	Inpatients	81 (26-198)	29% (342/1172)	Yes	Pulmonary samples; Clinically relevant extrapulmonary samples. No study defined algorithm.	No	Adults (>18 years); Multisite; LAM arm of a randomised controlled trial.
HIV positive adults irrespective of signs and symptoms of TB								
Bjerrum 2015	**Unselected;** 91% symptomatic	Outpatients (85%); Inpatients (15%)	127 (35-256)	12% (55/469)	No	Pulmonary samples	No	Adults (>18 years); Majority symptomatic.
Drain 2015	**Unselected;** proportion	Outpatient	248	17% (54/320)	No	Pulmonary samples	No	Adults (>18 years)

Study	Participants (% symptomatic)	Setting	Median CD4 cell count per µL (IQR)	TB prevalence % (n/N)	Did the study avoid inappropriate exclusion	Specimens collected	High quality reference standard*	Unique Study Characteristics
	symptomatic not stated		(107-379)					
Floridia 2017	**Unselected;** 34% symptomatic	Outpatient	278 (142-395)	9% (90/972)	No	Pulmonary samples	No	Adults (> 15 years). LAM guided treatment.
Hanifa 2016	**Unselected;** 53% symptomatic	Outpatient	111 (56-161)	9% (40/408)	Yes	Pulmonary samples; Blood culture for all	Yes	Adults (>18 years); CD4 < 200; Reference standard included any sample taken within six months from enrolment.
LaCourse 2016	**Unselected;** 19% symptomatic	Outpatient	437 (342-565)	1% (3/266)	No	Pulmonary samples	No	Pregnant women (>16 years) attending ANC; Healthy population; one person with CD4< 400; Few TB cases (n=3).
Lawn 2017	**Unselected;** 91% symptomatic	Inpatients	149 (55-312)	33% (139/413)	Yes	Pulmonary samples; Blood culture for all; Clinically relevant extrapulmonary samples	Yes	Adults (>18 years). Included many samples from different sites
Thit 2017	**Unselected;** 33% symptomatic	Outpatients (90%); Inpatients (10%)	270 (128-443)	10% (54/517)	Yes	Pulmonary samples	No	Adults (median age 34). Reference standard included samples taken within six months from enrolment.

Abbreviations: TB: tuberculosis; AlereLAM: AlereLAM: Alere Determine™ TB lipoarabinomannan assay; Xpert: Xpert MTB/RIF

* For a microbiological reference standard, we considered a higher quality reference standard to be one in which two or more specimen types were evaluated for TB diagnosis in all participants as part of a defined standardized study algorithm.

**Huerga 2017 excluded participants from analysis if missing Xpert results or culture contaminated for any of the samples in the absence of a positive result; overall samples size 474 (156 with TB); 275 included in analysis (156 with TB).

*** Pandie 2016 excluded a large number of non-TB participants from analysis; Overall samples size 102 (36 with TB); 38 included for analysis (36 TB cases).

Appendix 9: Diagnostic accuracy of AlereLAM among HIV-positive children, summary

Background

Among children in 2014, there were an estimated one million incident TB cases and 140,000 deaths attributable to TB. Approximately 40% of TB deaths were among those coinfected with TB and HIV (Carlucci 2017).

Diagnosis of TB in children

Conventional diagnosis by culture or microscopy yields a positive result in less than 50% of children with clinically diagnosed TB and HIV infection (Thomas 2016). Active TB, therefore, remains unrecognised in a large number of children in high burden countries as evident from autopsy studies from five African countries that identified TB in roughly 10% of 811 children (both HIV-positive and HIV-negative) who died from presumed pneumonia (Bates 2013).

To obtain specimens for microscopy or culture, methods used in children include gastric lavage (GL), which requires uncomfortable insertion of a nasogastric tube and induced sputum. Sputum induction requires special facilities (negative pressure) for infection control and nebulization equipment driven by high flow air or oxygen, not available in rural areas of low-income countries. Using sputum induction for children with presumed pulmonary TB, a study conducted in South Africa reported microbiologic confirmation in 11% of cases (Connell 2011). A lower yield (3.8% to 7%) has been reported for nasopharyngeal aspirate. Bronchoalveolar lavage (BAL) is a resource-intensive and invasive procedure that has a lower yield for culture, compared with GL; therefore, BAL is not indicated for microbiologic confirmation of TB in children.

The WHO recommends the use of Xpert MTB/RIF (Xpert) as follows.

1) Xpert should be used rather than conventional microscopy and culture as the initial diagnostic test in children suspected of having MDR-TB or HIV-associated TB (strong recommendation, very low-certainty evidence)

2) Xpert may be used rather than conventional microscopy and culture as the initial test in all children suspected of having TB (Conditional recommendation acknowledging resource implications, very low-certainty evidence)

3) Xpert may be used as a replacement test for usual practice (including conventional microscopy, culture, and/or histopathology) for testing of specific non-respiratory specimens (lymph nodes and other tissues) from children suspected of having extrapulmonary TB (conditional recommendation, very low-certainty evidence) and

4) Xpert should be used in preference to conventional microscopy and culture as the initial diagnostic test in testing cerebrospinal fluid specimens from children suspected of having TB meningitis (strong recommendation given the urgency of rapid diagnosis, very low-certainty evidence) (WHO 2014).

The next-generation assay, Xpert MTB/RIF Ultra (Ultra), has shown improved sensitivity for detection of TB in HIV-positive people. The WHO now recommends Ultra as a replacement for the current Xpert MTB/RIF cartridge (WHO Ultra 2017).

Methods

We performed literature searches up to 11 May 2018 as part of a larger search for studies in adults with the same inclusion criteria except for age. We included studies that evaluated Alere Determine™ TB LAM Ag test (AlereLAM) on urine specimens. The target condition was active TB disease, which includes pulmonary and extrapulmonary TB. Age groups were defined as younger children ≤ 5 years; adolescents, 10 to 19 years; and older children, 6 to 19 years.

Two review authors independently extracted data on methodological quality and 2x2 values for AlereLAM for TB against a microbiological reference standard. Given the differences in population and setting, we did not perform meta-analyses and provide sensitivity and specificity estimates for individual studies.

Results

We identified three published studies involving 266 HIV-positive children that evaluated the accuracy of AlereLAM for TB as the result of a broader search for studies in adults and children using the same inclusion criteria (Kroidl 2015; LaCourse 2018; Nicol 2014). All three studies took place in high TB/HIV burden countries in Africa: Kroidl 2015 in Tanzania; LaCourse 2018 in Kenya; and Nicol 2014 in South Africa.

Methodological quality of included studies

In the Patient Selection Domain, we considered two studies (67%) to have low risk of bias because the study used consecutive or random enrolment of participants and avoided inappropriate exclusions (Kroidl 2015; LaCourse 2018). We considered one study to have high risk of bias because children who could not produce sputum despite sputum induction were excluded (Nicol 2014). In the Index Test we considered all three studies to have low risk of bias. In the Reference Standard Domain, we considered two studies (67%) to have low risk of bias and one study to have high risk of bias because we thought the reference standard used was unlikely to correctly classify the target condition (Nicol 2014). In the Flow and Timing Domain we considered one study to have high risk of bias because not all patients received the same reference standard (Kroidl 2015). Applicability in all domains was of low concern in all three studies, (Figure A1 and Figure A2).

Figure A1. Risk of bias and applicability concerns graph.
Review authors' judgements about each domain presented as percentages across included studies in children.

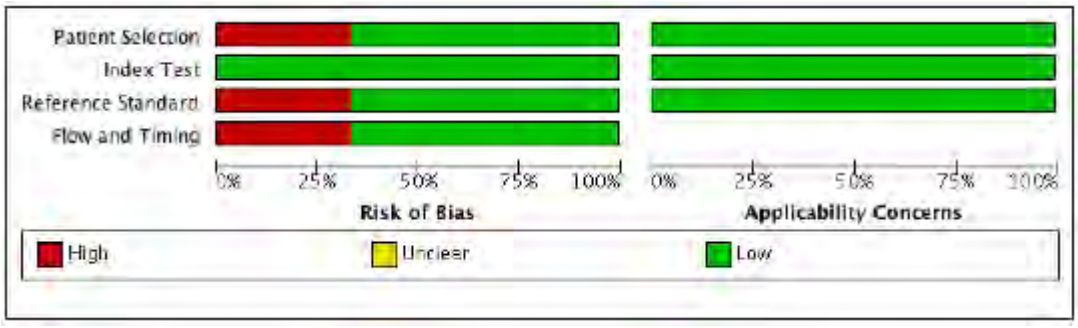
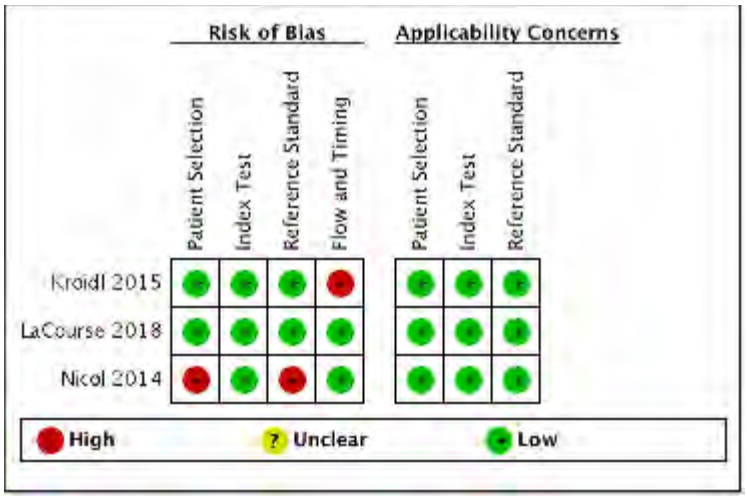


Figure A2. Risk of bias and applicability concerns summary.
Review authors' judgements about each domain for each included study in children.



Findings
Kroidl 2015 enrolled children six weeks to 14 years, median age (interquartile range (IQR)) 6.8 years (3.9 to 9.5) for all participants, including HIV-positive and HIV-negative children. LaCourse 2018 enrolled children aged 12 years or less, median age (IQR) 24 months (13 to 58). Nicol 2014 enrolled children aged 15 years or less, median age (IQR) 42.5 months (19.1 to 66.3) for all participants, including HIV-positive and HIV-negative children.

Kroidl 2015 and Nicol 2014 involved HIV-positive children with TB symptoms. LaCourse 2018 involved HIV-positive children hospitalized for acute illness irrespective of TB signs and symptoms. Kroidl 2015 was conducted in an outpatient setting, LaCourse 2018 in an inpatient setting, and Nicol 2014 in both an inpatient and an outpatient setting. The prevalence of microbiologically-confirmed TB in the studies was 40% in Kroidl 2015, 7% in LaCourse 2018, and 22% in Nicol 2014. Regarding immunosuppression, in Kroidl 2015, 65% of children had advanced or severe immunosuppression; in LaCourse 2018, 70% of children had severe immunosuppression; and in Nicol 2014, 53% of children had advanced or severe immunosuppression. See Table.

Table. Characteristics of Included Studies, Children

Study	Country	Age of enrolment	Presence of TB symptoms?	Setting	TB prevalence % (n/N)	Percent children with advanced or severe immunosuppression
Kroidl 2015	Tanzania	Six weeks to 14 years, median 8.8 years (IQR 3.9 to 9.5)*	TB symptoms	Outpatient	40% (12/30)	65%
LaCourse 2018	Kenya	≤ 12 years, median 24 months (IQR 13 to 58)	Irrespective of TB symptoms	Inpatient	7% (9/130)	70%
Nicol 2014	South Africa	≤ 15 years, median 42.5 months (IQR 19.1 to 66.3)*	TB symptoms	Inpatient and outpatient	22% (23/160)	53%

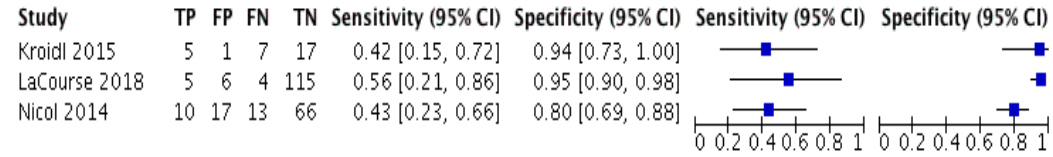
* For all participants, including HIV-positive and HIV-negative children.
IQR: Interquartile range

We first present a summary and then present findings for the specific PICO questions.

AlereLAM testing, studies in children with TB symptoms and irrespective of TB signs and symptoms

All settings
We identified three studies involving 266 children (Kroidl 2015; LaCourse 2018; Nicol 2014), ***Figure A3***. In all settings, including all children, sensitivity and specificity (95% CI) were 42% (15% to 72%) and 94% (73% to 100%), (30 participants, outpatient) Kroidl 2015; 56% (21% to 86%) and 95% (90% to 98%), (130 participants, inpatient) LaCourse 2018; and 43% (23% to 66%) and 80% (69% to 88%), (106 participants, both inpatient and outpatient) Nicol 2014.

Figure A3. Forest plots of AlereLAM sensitivity and specificity for TB in HIV-positive children with TB symptoms and irrespective of TB signs and symptoms, all settings.

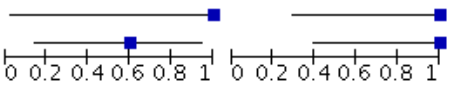


By age group
Kroidl 2015 and LaCourse 2018 provided AlereLAM accuracy data by age group, ***Figure A4***. Stratified by age group, in adolescents, AlereLAM sensitivities were 100% (33% to 100%) (four participants, inpatient) LaCourse 2018, and 60% (15% to 95%) (nine participants, outpatient) Kroidl 2015; in both studies, specificity was 100%. In children ≤ 5 years, sensitivities were 50% (7% to 93%) (95 participants, inpatient) LaCourse 2018, and 25% (1% to 81%) (13 participants, outpatient) Kroidl 2015; corresponding specificities were 93% (86% to 98%) and 89% (52% to 100%).

Figure A4. Forest plots of AlereLAM sensitivity and specificity for TB in HIV-positive children with TB symptoms and irrespective of TB symptoms, by age group.

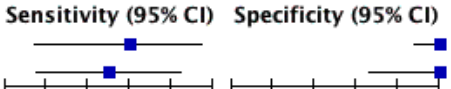
Adolescents

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
LaCourse 2018	1	0	0	3	1.00 [0.03, 1.00]	1.00 [0.29, 1.00]
Kroidl 2015	3	0	2	4	0.60 [0.15, 0.95]	1.00 [0.40, 1.00]



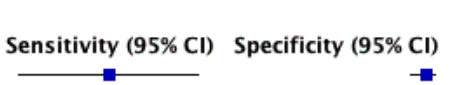
Older children

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
LaCourse 2018	3	0	2	29	0.60 [0.15, 0.95]	1.00 [0.88, 1.00]
Kroidl 2015	4	0	4	9	0.50 [0.16, 0.84]	1.00 [0.66, 1.00]



Children ≤5

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
LaCourse 2018	2	6	2	85	0.50 [0.07, 0.93]	0.93 [0.86, 0.98]
Kroidl 2015	1	1	3	8	0.25 [0.01, 0.81]	0.89 [0.52, 1.00]



PICO 1. What is the diagnostic accuracy of AlereLAM for the diagnosis of TB in all HIV-positive children with signs and symptoms of TB?

1.a.ii. Inpatient settings, adolescents
We did not identify any studies addressing this question.

1.a.iii. Inpatient settings, older children
We did not identify any studies addressing this question.

1.b.ii. Outpatient settings, adolescents
We identified one study involving nine HIV-positive children, five (56%) with TB, Kroidl 2015. Sensitivity and specificity (95% CI) were 60% (15, 95) and 100% (40,100).

1.b.iii. Outpatient settings, older children
We identified one study involving 17 HIV-positive children, eight (47%) with TB, Kroidl 2015. Sensitivity and specificity (95% CI) were 50% (16, 84) and 100% (66,100).

1.c.ii. All settings, adolescents
We identified one study involving nine HIV-positive children, five (56%) with TB, Kroidl 2015. The data are from an outpatient setting. Sensitivity and specificity (95% CI) were 60% (15, 95) and 100% (40, 100).

1.c.iii. All settings, older children
We identified one study involving 17 HIV-positive children, eight (47%) with TB, Kroidl 2015. The data are from an outpatient setting. Sensitivity and specificity (95% CI) were 50% (16, 84) and 100% (66, 100).

1.d. Inpatient settings, children ≤ 5 years
We did not identify any studies addressing this question.

1.e. Outpatient settings, children ≤ 5 years
We identified one study involving 13 HIV-positive children, four (31%) with TB, Kroidl 2015 Sensitivity and specificity (95% CI) were 25% (1, 81) and 89% (52,100).

1.f. All settings, children ≤ 5 years

We identified one study involving 13 HIV-positive children, four (31%) with TB, Kroidl 2015. The data are from an outpatient setting. Sensitivity and specificity (95% CI) were 25% (1, 81) and 89% (52, 100).

PICO 2. What is the diagnostic accuracy of AlereLAM for the diagnosis of TB in all HIV-positive children irrespective of signs and symptoms of TB?

2.a.ii. Inpatient settings, adolescents

We identified one study involving four HIV-positive children, one (25%) with TB, LaCourse 2018. Sensitivity and specificity (95% CI) were 100% (3, 100) and 100% (29, 100).

2.a.iii. Inpatient settings, older children

We identified one study involving 34 HIV-positive children, five (15%) with TB, LaCourse 2018. Sensitivity and specificity (95% CI) were 60% (15, 95) and 100% (88, 100).

2.b.ii. Outpatient settings, adolescents

We did not identify any studies addressing this question.

2.b.iii. Outpatient settings, older children

We did not identify any studies addressing this question.

2.c.ii. All settings, adolescents

We identified one study involving four HIV-positive children, one (25%) with TB, LaCourse 2018. The data are from an inpatient setting. Sensitivity and specificity (95% CI) were 100% (3, 100) and 100% (29, 100).

2.c.iii. All settings, older children

We identified one study involving 34 HIV-positive children, five (15%) with TB, LaCourse 2018. The data are from an inpatient setting. Sensitivity and specificity (95% CI) were 60% (15, 95) and 100% (88, 100).

2.d. Inpatient settings, children ≤ 5 years

We identified one study involving 95 HIV-positive children, four (4%) with TB, LaCourse 2018. Sensitivity and specificity (95% CI) were 50% (7, 93) and 93% (86, 98).

2.e. Outpatient settings, children ≤ 5 years

We did not identify any studies addressing this question.

2.f. All settings, children ≤ 5 years

We identified one study involving 95 HIV-positive children, four (4%) with TB, LaCourse 2018. The data are from an inpatient setting. Sensitivity and specificity (95% CI) were 50% (7, 93) and 93% (86, 98).

Discussion

This systematic review on the urine lateral flow lipoarabinomannan assay, AlereLAM, for active TB in children living with HIV summarizes the current literature and includes three studies. As the studies enrolled children aged 15 years and less and younger children (median age, range 24 months to 6.8 years), the results may not be applicable to older children. All studies took place in high TB/HIV burden countries in Africa. We corresponded with two study authors (Kroidl and LaCourse) to ensure that we had accurate data for AlereLAM applied using the current manufacturer's instructions. In individual studies, AlereLAM sensitivities were 56%, 42%, and 43%; corresponding specificities were 94%, 95%, and 80%.

AlereLAM specificity was lower in children ≤ 5 years than in older children, based on limited data. Urine collection was noted to be difficult in younger and sicker children. In addition, urine collection in children may be affected by dehydration or other medical problems.

Authors' conclusions

We found limited evidence on the accuracy of AlereLAM in children living with HIV. There were too few studies and participants to draw conclusions.

References for included studies

Kroidl 2015
Kroidl I, Clowes P, Reither K, Mtafya B, Rojas-Ponce G, Ntinginya EN, et al. Performance of urine lipoarabinomannan assays for paediatric tuberculosis in Tanzania. European Respiratory Journal 2015;46(3):761-70.

LaCourse 2018
LaCourse SM, Pavlinac PB, Cranmer LM, Njuguna IN, Mugo C, Gatimu J, et al. Stool Xpert MTB/RIF and urine lipoarabinomannan for the diagnosis of tuberculosis in hospitalized HIV-infected children. AIDS 2018;32(1):69-78.

Nicol 2014
Nicol MP, Allen V, Workman L, Isaacs W, Munro J, Pienaar S, et al. Urine lipoarabinomannan testing for diagnosis of pulmonary tuberculosis in children: a prospective study. Lancet Global Health 2014;2(5):e278–e284.

Additional references

Bates 2013
Bates M, Mudenda V, Mwaba P, Zumla A. Deaths due to respiratory tract infections in Africa: a review of autopsy studies. Current Opinion Pulmonary Medicine 2013;19(3):229-37.

Carlucci 2017
Carlucci JG, Blevins PM, Kipp AM, Lindegren ML, Du QT, Renner L, et al. Tuberculosis treatment outcomes among HIV/TB-coinfected children in the International, Epidemiology Databases to Evaluate AIDS (IeDEA) Network. Journal of Acquired Immune Deficiency Syndrome 2017;75(2):156-163.

Connell 2011
Connell TG, Zar HJ, Nicol MP. Advances in the diagnosis of pulmonary tuberculosis in HIV-infected and HIV-uninfected children. The Journal of Infectious Diseases 2011;204: 2011;204(Suppl 4):S1151–58.

Thomas 2016
Thomas TA. Challenges in diagnosing childhood tuberculosis. Infectious Disease Journal 2016;25(2):26.

WHO 2014
World Health Organization. Guidance for national tuberculosis programmes on the management of tuberculosis in children. apps.who.int/medicinedocs/documents/s21535en/s21535en.pdf 2014.

WHO Ultra 2017
World Health Organization. WHO meeting report of a technical expert consultation: non-inferiority analysis of Xpert MTB/RIF Ultra compared to Xpert MTB/RIF. WHO/HTM/TB/2017.04. Geneva: WHO 2017.

Web Annex D.13. Economic evaluations of LF-LAM for the diagnosis of active tuberculosis in HIV-positive individuals: an updated systematic review

ACKNOWLEDGEMENTS

Alice Zwerling

Secondary reviewers: Dr Sazia Naz and Olivia D'Silva

School of Epidemiology and Public Health
University of Ottawa
Ottawa, Canada

Executive Summary

We carried out a systematic review of economic evaluations on the urine-based lateral flow lipoarabinomannan assay AlereLAM (Alere Determine ™ TB LAM Ag, Abbott, Palatine, IL, USA, previous Alere Inc., Waltham, MA, USA) for the diagnosis of active TB in HIV-positive individuals. The objective of this review was to summarize current evidence and understand the costs, cost-effectiveness and affordability (in terms of budget impact) of AlereLAM implementation for diagnosis of tuberculosis (TB) among HIV-positive populations. We identified 6 studies all from settings in sub-Saharan Africa. Study methods and populations were heterogeneous, assessing a range of diagnostic algorithms, and only 4 studies assessed cost-effectiveness.

Economic evidence for the implementation and scale-up of AlereLAM is still limited. Existing studies show a consistent trend, suggesting a high probability that AlereLAM could be cost-effective in a population of African adults living with HIV (particularly amongst hospitalized patients). However, with only a few studies and key differences in modeling approaches, assumptions, diagnostic algorithms assessed, analytical techniques, and study settings, generalizability and more specifically applicability to other settings is limited.

Inclusion of costs associated with antiretroviral therapy and HIV care resulted in higher incremental cost-effectiveness ratios as TB diagnostic costs represented just a small proportion of total increased costs when HIV care is included. Models found cost-effectiveness of AlereLAM to be robust across a variety of sensitivity analyses, variations in key parameters and across different country settings and scenarios. Key parameters that are likely influential on cost-effectiveness include: TB prevalence, target population, and AlereLAM specificity, cost of treating TB and HIV and life expectancy post TB survival, and time horizon. However, one detailed micro-costing study published in 2018 estimates unit test costs for AlereLAM implementation several fold higher (US$ 23) than most current models ($2-4). Underestimation of AlereLAM unit costs could result in overly optimistic cost-effectiveness profiles.

While current evidence is consistent in suggesting AlereLAM is likely cost-effective among HIV-positive patients in sub-Saharan Africa, caution should be used when extrapolating from a small number of studies, and additional evidence from a wider range of populations, settings and diagnostic approaches will be necessary.

Background

Mortality due to tuberculosis (TB) remains high among persons living with HIV, with TB accounting for 32% of AIDS related deaths in 2017[1]. Diagnosis of tuberculosis (TB) among HIV-positive populations remains a critical challenge in the fight against TB, conventional diagnostics are not as effective in this population due to an inability to provide sputum required for most standard TB diagnostics[1]. In 2015, the WHO issued conditional recommendations endorsing a urine lipoarabinomannan (urine LAM) assay (Determine TB-LAM; Alere, MA, USA) for use in adult inpatients with HIV and TB symptoms and who have CD4 count <100 cells per μL or who are seriously ill[2,3].

Lipoarabinomannan (LAM) is a glycolipid found in the outer cell wall of mycobacteria. During active TB disease, LAM antigen is released from metabolically active or degrading bacterial cells and passes into the urine. The first commercially available test for urine LAM used an enzyme linked immunosorbent assay (ELISA) format (Clearview TB ELISA, Alere Inc, Waltham, MA, USA). Later developments resulted in a lateral flow urine lipoarabinomannan (LF-LAM) assay, which is commercially available to detect active TB (Alere DetermineTM TB LAM Ag, Alere Inc, Waltham, MA, USA). The test can be performed at the point of care, and does not require technical expertise. A 60 μL sample of urine is applied to a nitrocellulose test strip and incubated for 25 minutes. The result can then be easily read without additional equipment by comparing band intensity with a manufacturer-supplied reference card, resulting in a positive/negative interpretation. Its diagnostic sensitivity has been shown to be improved amongst HIV-positive patients with lower CD4 counts. AlereLAM has many logistical advantages, including its ease of use, lack of requirements for equipment and infrastructure, ability to be performed at the point of care, with results in less than a half hour and low cost (initially marketed at US$ 3.50USD per test).

AlereLAM has shown great potential for use in HIV-positive persons, but the economic evidence to date has been limited. A 2015 systematic review tittled: **"A Systematic Review of Economic Evaluations of the Lateral Flow Urine Lipoarabinomannan Assay for Diagnosis of Active Tuberculosis in HIV-infected Individuals"** prepared by Hanrahan and Dowdy, identified only 2 eligible studies3. Understanding the costs, cost-effectiveness and affordability of AlereLAM in HIV-positive individuals can provide important evidence for policy makers needing to make decisions around scale-up of the test in TB and other programmes that care for HIV-positive populations.

Objective

To perform an updated systematic review of the published literature on economic evaluations on the urine based lateral flow lipoarabinomannan assay AlereLAM (Alere Determine ™ TB LAM Ag, Abbott, Palatine, IL, USA, previous Alere Inc., Waltham, MA, USA) for the diagnosis of active TB in HIV-positive individuals. To summarize current evidence and further understand the costs, cost-effectiveness and affordability of AlereLAM implementation for TB diagnosis among HIV-positive populations. Affordability was considered with respect to budget impact assessments performed in specific countries under given scenarios/conditions.

Methods

Types of studies considered

Studies were included if they evaluated the index test, AlereLAM, for the detection of active TB disease among HIV-positive individuals, and included an economic evaluation in the analysis. Our search term outlined below, were designed to broadly capture any economic evaluations or studies mentioning cost or disability-adjusted life years (DALYs) or quality-adjusted life years (QALYs) and was not limited to cost-effectiveness analyses.

Examples of eligible economic studies included cost-minimization analysis, cost-consequence analysis, cost-effectiveness analysis, and cost-benefit analysis. We considered studies eligible that used either primary or secondary data sources (i.e., published literature) for either economic or epidemiologic parameters. Studies were excluded if they reported only the cost of the AlereLAM test without including health systems delivery costs or if there was no link to health outcomes such as incremental yield, mortality, or DALYs/QALYs. Only studies published in English were included.

Search strategy & Data Sources

Using the same search strategy as the 2015 report, we performed a search of three online databases: PubMed, EMBASE and Web of Science for new studies published from January 1, 2010 through September 31, 2018. We reviewed citations of all eligible articles, guidelines and reviews for additional studies.

The search strategy used was the following: (tuberculosis OR TB) AND (lipoarabinomannan OR LAM) AND (test OR assay OR antigen OR Ag OR lateral flow assay OR urine antigen OR point of care) AND (((((cost-benefit) OR (cost) OR (economic) OR (cost effectiveness) OR (cost-utility) OR (disability adjusted life year) OR DALY OR (quality-adjusted life year) OR QALY OR (cost benefit analysis) OR (cost effectiveness analysis))) OR ((quality of life) OR (utility))) AND (HIV OR Human Immunodeficiency Virus).

Search terms were adapted slightly for each database as required.

Study selection

The study selection followed PRISMA guidelines.[4,5] Potentially relevant studies were identified through electronic searches of the online databases as described above, and duplicates were removed. An initial abstract review of each study was completed independently by two reviewers; articles were excluded if they did not evaluate AlereLAM or if they were reviews, letters or opinion pieces (i.e. no original data). Full text review was then completed on remaining articles, and articles that met predetermined inclusion criteria (evaluated AlereLAM for active TB disease and included an economic evaluation, estimated costs beyond the price of the assay and or linked those costs to health outcomes), were retained for the review. In an effort to include and review all available economic data for AlereLAM, studies were not restricted to cost-effectiveness and cost-utility analyses.

Data extraction

Full texts of included studies were independently reviewed by two reviewers, including published supplemental material, with all disagreements resolved by consensus and discussions with a third reviewer. Assessment of the quality of each economic evaluation was guided by the Consensus Health Economic Criteria (CHEC) list.[6,7]

The study design data elements extracted from each study included: the primary research question, country setting, year of study, patient population, clinical setting, AlereLAM diagnostic scenarios, comparison diagnostic scenarios, analysis perspective, analytic time horizon, type of economic evaluation, source of costing, primary outcome measure, secondary outcome measure, type of model, types of sensitivity and uncertainty analyses performed and willingness-to-pay threshold.

Key model parameters were extracted and presented in tables, including epidemiologic, diagnostic testing and treatment and outcome parameters. Key outcomes for each study were extracted including: cost per patient for each diagnostic strategy, and incremental cost effectiveness ratio (ICER) per effectiveness or utility measure (e.g. DALY averted). Costs were presented in USD (United States Dollars).

Results

Study selection

A search of online databases as of September 31, 2018 returned a total of 130 articles; the search was updated as of February 1, 2019 to include a recently published paper. Duplicates were excluded (n=33) for a total of 98 articles that were screened and reviewed for study eligibility. Articles were excluded if they did not include data on AlereLAM (n=27) and if they did not include original data or analysis, (i.e. review, letters, opinion, n=20). A full text review was performed on the remaining 51 articles, and those that did not include an economic evaluation were excluded (n=45), leaving 6 articles eligible for inclusion in our review[8-13].

Study characteristics

Characteristics of the 6 included studies are summarized in Table 1. All studies were based in settings in sub-Saharan Africa, primarily South Africa along with Uganda, Mozambique and Malawi. It is worthwhile noting that 4 of the 6 studies included were only recently published, in 2018 and 2019. Three studies focused on hospitalized HIV-positive patients exclusively, two studies on outpatients living with HIV, and one study examined both inpatient and outpatient settings. With the exception of the recent Reddy 2019 cost-effectiveness analysis which assessed using AlereLAM testing protocols irrespective of CD4 cell count or TB symptoms, studies assessed implementation of AlereLAM containing algorithms only in targeted subgroups (CD4 cell count <100, <150 and <200 and/or among those with TB symptoms).

A variety of AlereLAM diagnostic strategies and reference strategies were compared. Sun 2013 compared the addition of AlereLAM to a combination reference strategy: sputum smear microscopy (SSM) or Xpert MTB/RIF (Xpert) along with clinical judgement. Sun 2013 also explored an alternative scenario where Xpert replaced SSM in the combined reference strategy. Shah 2013 assessed the addition of AlereLAM to two reference strategies: either SSM or Xpert alone. Boyles 2018 assessed several different testing algorithms where subsequent tests were performed only in the case of a negative precedent test result. Algorithms assessed included Xpert followed by culture, AlereLAM, followed by Xpert, AlereLAM followed by Xpert then culture, AlereLAM followed by Xpert SI (Sputum induction), Xpert SI followed by culture, AlereLAM followed by Xpert SI then culture. In the costing analysis Mukora 2018 assessed the costs of AlereLAM and point of care haemoglobin testing in settings where same-day Xpert and smear microscopy results are not available. Orlando 2018 used a reference strategy including a 4 symptom screen for TB and then SSM for any with positive symptoms. This reference was compared with an Xpert for all participants approach and an AlereLAM strategy where AlereLAM was performed in all patients with CD4 counts <200, and Xpert performed in patients with CD4 counts >200 and in those with negative AlereLAM. Reddy 2019 used an Xpert for all reference strategy, comparing this with either Xpert + AlereLAM +urine Xpert or Xpert and AlereLAM. Reddy 2019 also assessed the cost-effectiveness of these interventions among a subgroup with CD4 cell count <100.

Three studies included cost-utility analyses and produced incremental cost-effectiveness ratios (ICER) (cost/DALY averted) as their primary outcome. Reddy 2019 performed a cost-effectiveness analysis with incremental cost per year of life saved as the primary outcome. Reddy 2019 was the first to include a budget impact assessment along with the cost-effectiveness analysis. Boyles 2018 calculated incremental yield and cost per patient associated with various diagnostic algorithms assessed, but did not further associate this with incremental health benefit or DALY averted. Mukora 2018 performed a detailed cost estimation of AlereLAM for use in South Africa, where the primary outcome was unit test cost with a focus on understanding clinic level and above clinic level unit costs. All studies employed health system perspective approaches. Sun 2013 explicitly limited their analysis to the TB program perspective and did not include downstream costs associated with antiretroviral therapy (ART). Only Shah 2013 and Reddy 2019 included costs associated with HIV care and ART.

The majority (4/6) of studies employed empirical data collection for at least some of the costing sources, while Sun 2013 and Boyles 2018 used exclusively published literature for costing sources. Shah 2013, Mukora 2018, Orlando 2018 and Reddy 2019 all used a combination of direct observation, timesheets, review of financial, study/program, laboratory and hospital records to build TB unit costs for their analysis. Sun 2013, Shah 2013 and Orlando 2018 all employed a static decision analysis model, while Reddy 2019 employed a modified cost-effectiveness of preventing AIDS complications-International (CEPAC-I) microsimulation model.

One-way sensitivity analyses were performed across all of the 4 cost-utility/cost-effectiveness studies. Sun 2013 and Shah 2013 also performed probabilistic sensitivity analyses while Reddy 2019 performed multiway deterministic sensitivity analyses across key parameters and scenarios. Shah 2013 and Sun 2013 both employed national per capital GDP as the suggested willingness-to-pay (WTP) threshold, while Orlando 2018 employed the three fold national GDP threshold often cited. Reddy 2019 took an alternative approach, suggesting the ICER of second line ART as determined by the CEPAC-1 model for each country under investigation. Input parameters for included studies are listed in Table 2.

The 6 included studies were assessed using the quality of health economics studies instrument assessment and the CHEERS checklist[6]. Results are summarized in Appendix Table 1. All 6 studies clearly described key methods, objectives, and alternatives compared. Appropriate analyses were performed where indicated in respective studies' objectives and design. Two of the 6 articles did not discuss potential conflict of interests, a further 2 studies did not discuss ethical and distributional issues.

Study findings

Impact of AlereLAM on TB case finding

Five included studies estimated impact of AlereLAM diagnostic strategies on TB case finding. Mukora 2018 performed a costing study only and did not report on potential impact of AlereLAM on TB case finding or cost-effectiveness.

In a prospective cohort study of hospitalized HIV-positive patients with TB symptoms Boyles 2018 assessed the incremental yield of various diagnostic algorithms containing AlereLAM. Yields ranged from 24.3% in the Xpert/culture approach to 50.9% in the AlereLAM/Xpert approach, 52.1% in the AlereLAM/Xpert/culture, 92.3% in the AlereLAM/Xpert SI, 95% in Xpert SI/culture, and 95.9% in AlereLAM/Xpert SI/Culture approach. AlereLAM sensitivity in this cohort was 35.5%, significantly lower than parameter values used in Sun 2013, 66% and Shah 2013, 49%. Incremental yield of AlereLAM strategies was also much lower (1.2- 3.6%) compared with Sun 2013 and Shah 2013, due to the use of culture as the comparison.

Orlando 2018 compared an Xpert for all approach and an AlereLAM/Xpert algorithm with the standard of care (4 symptom screen and smear microscopy for those with symptoms) in a simulated cohort of 1000 HIV-positive outpatients initiating ART in Mozambique. An estimated 1281 and 1254 DALYs were saved using the Xpert and AlereLAM/Xpert approach respectively, compared with standard of care which averted 1107 DALYs, representing approximately a 13% incremental yield in DALYs averted using AlereLAM/Xpert compared with standard of care containing SSM, consistent with incremental yield employed by earlier studies.

The most recently published study, Reddy 2019 used the CEPAC-I model to simulate STAMP trial results and project long term outcomes[14]. Reddy 2019 adapted the CEPAC-I model, a validated microsimulation model of HIV related disease and treatment, to incorporate TB natural history, diagnosis, and treatment. The primary intervention assessed included sputum Xpert, AlereLAM and urine Xpert among unselected

hospitalized patients with HIV. When the model was calibrated to STAMP trial outcomes, an estimated absolute increase in diagnostic yield of 55% was seen in Malawi and 23% in South Africa using the AlereLAM intervention compared with standard of care (Xpert), with an estimated increase in life expectancy of 1.2 years in Malawi and 0.5 years in South Africa. Differences in yield between Malawi and South Africa were driven by lower probability of obtaining a sputum sample in Malawi and higher rates of empirical treatment in South Africa. It is worthwhile noting that the STAMP trial found only weak evidence (p=0.07) of a 2.8% reduction in all-cause mortality at 2 months with the AlereLAM intervention compared with standard of care. Were this mortality reduction not to hold true, estimates of impact and cost-effectiveness would be impacted.

Employing hypothetical cohorts, Sun 2013 estimate AlereLAM + SSM + existing diagnostics detects 90% of TB cases compared with 73% using the standard of care (SSM + existing diagnostics) when used in a hospitalized population. Shah 2013 estimated AlereLAM + SSM could detect 81% (95%Uncertainty Range (UR) 62-91%) of TB cases in a hypothetical Ugandan cohort (inpatients and outpatients) versus 66% (95%UR 41-80) using SSM alone. While Sun 2013 had a higher yield using their standard of care approach (as you might expect among a hospitalized population), both studies estimated a 15-17% incremental yield with the addition of AlereLAM compared to standard of care without Xpert. Sun 2013 did find that the addition of AlereLAM increased false positives by 19% resulting in increased unnecessary TB treatment. Shah 2013 assessed the addition of AlereLAM to Xpert strategy, resulting in a case detection of 93% (95%UR 81-96%) of TB cases, compared to 87% (95%UR 41-80%) with Xpert alone.

Costs of ALereLAM for TB detection & budget impact assessment

From the Mozambique setting, Orlando 2018 estimated total costs of three approaches. Orlando 2018 included cost of diagnostics, and treatment, but also cost of newly transmitted infections due to delayed diagnosis and costs of additional transmitted infections due to health system delay. When considering only diagnostic and treatment costs, standard of care (SOC) was much less expensive at total cost of US$ 5,893 (US$ 5.89/person screened) compared with US$ 15,731 (US$ 15.73/person screened) with Xpert and US$ 16,522 (US$ 16.52/ per person screened) with Alere/Xpert interventions. However, when cost of additional infections due to health system delay were considered, standard of care became several fold more costly (US$ 87,519 & US$ 147,226 SOC vs US$ 18,168 & $92,263 Xpert vs US$ 18,959 & $113,196 AlereLAM).

Reddy 2019 estimated discounted per person lifetime health-care costs, therefore including TB diagnostics but also costs associated with TB treatment, ART and HIV care, resulting in higher costs per person screened. Reddy 2019 found per person lifetime health-care costs of US$ 3,450 using the standard of care approach and US$ 3,790 in the intervention approach for Malawi and US$ 8,500 and US$ 8,770 in South Africa.

Reddy 2019 also performed a budget impact assessment, providing some evidence on potential affordability of these approaches within the local health expenditures budget. They estimated the increased cumulative health-care expenditures among screened individuals due to diagnostic test costs alone (ALereLAM and urine Xpert) to be US$ 10 million (11.2%) over 2 years and $37 million over 5 years (10.8%) in Malawi and US$ 73 million (2.4%) over 2 years and US$ 261 million (2.8%) over 5 years in South Africa. Reddy 2019 demonstrate TB diagnostics represent a small percentage of total health care expenditures, which are driven largely by cost of ART and non-ART HIV care including hospitalizations.

While Mukora 2018 did not assess diagnostic yield of AlereLAM, they investigated the full economic costs in a South African setting (from the health care system perspective) of introducing AlereLAM as an initial TB test among HIV-positive outpatients with CD4 count <150 cell/μL. Mukora 2018 employed a detailed micro-costing approach including costs from both the clinic level and above clinic level, across non-governmental organizations (NGO) and department of health (DoH) implementers/clinics and included costs

from both start-up and implementation periods. Mukora 2018 estimated a total unit cost of AlereLAM testing at US$ 23.55 (NGO clinics) and $22.72 (department of health (DOH) operated clinics). Unit costs were higher than have been reported in other studies from South Africa largely driven by the inclusion of both clinic level (US$ 11.49 NGO & US$ 10.85 DOH) and above clinic level costs (US$ 12.06 NGO & US$ 11.87 DOH).

Using published costing data for South Africa limited to unit test cost, Boyles 2018 calculated the cost per patient for each algorithm. Cost per patient screened by each algorithm generally increased with increasing diagnostic yield and ranged from US$ 10.5 for Xpert/Culture and AlereLAM/Xpert, US$ 12.5 for the AlereLAM/Xpert/culture, US$ 37.2 for the AlereLAM/Xpert SI, US$ 49.6 for Xpert SI/culture, and US$ 42 for AlereLAM/Xpert SI/culture approach. Boyles 2018 did not perform a cost-effectiveness analysis or calculate incremental cost-effectiveness ratios.

Using non-empirical costing sources, Sun 2013 estimated total cost (including diagnosis and treatment) per patient evaluated for Uganda and South Africa. In South Africa, SSM plus existing diagnostics was estimated to cost US$ 243 /patient evaluated while the addition of AlereLAM resulted in a total cost of US$ 308/patient evaluated. In Uganda, standard of care cost US$ 71 and US$ 92 with the addition of AlereLAM.

Shah 2013 employed empirical costing for TB diagnostics in Uganda. For standard of care including SSM total per-patient costs were estimated at US$ 62 (95% UR US$ 37-116) and US$ 86 (95% UR US$ 57-137) for Xpert alone and US$ 91 (95% UR US$ 60-163) for Xpert plus LF-LAM, very similar to total costs estimated in Sun 2013. Differences were largely driven by different AlereLAM specificity used across the studies (95% versus 97%), resulting in higher treatment costs due to false-positive test results in Sun 2013. Shah 2013 were able to examine test component costs, diagnostic test cost alone for SSM was estimated to be US$ 15.16 per patient, while SSM + AlereLAM cost US$ 27.52, Xpert cost per test were estimated at US$ 31.80 and Xpert + AlereLAM was $36.55.

Shah 2013 demonstrated the major drivers of AlereLAM test costs are consumables (92%) with labor accounting for just 8% and equipment and overheads less than 0.1%. For SSM, consumables made up 59% of diagnostic costs, while equipment 20%, labor 16% and overhead 5%. For Xpert, consumables made up 70% of diagnostic costs, equipment 23%, labor 2% and overhead 5%.

Cost-effectiveness of TB diagnostic algorithms

Compared with the standard of care (smear microscopy for those positive on symptom screen), Orlando 2018 found the Xpert and Xpert/Alere approaches were highly cost-effective with ICERs of US$ 56.54/DALY averted for the Xpert approach and US$ 72.34/DALY averted for AlereLAM/Xpert. The smaller ICERs compared with earlier studies are driven partly by increased benefits associated with averted transmission and new infections. When cost of newly transmitted infections was included, Xpert and AlereLAM + Xpert approaches were cost-savings compared with the standard of care.

Using the modified CEPAC-I model calibrated to STAMP trial results, Reddy 2019 found Xpert + AlereLAM + urine Xpert to be cost-effective among unselected hospitalized HIV patients with ICERs of US$ 450/YLS (Years of life saved, YLS) in Malawi and US$ 840/YLS in South Africa compared with standard of care (Xpert alone). The modified intervention of Xpert + AlereLAM was even more cost-effective with ICERs of US$ 420/YLS in Malawi and US$ 810/YLS in South Africa compared with standard of care. Increased ICERs are due to inclusion of downstream costs associated with lifelong ART and HIV care.

Among the two cost-effectiveness studies included in the previous systematic review in 2015: Sun 2013 found that SSM + existing diagnostics + AlereLAM was highly cost-effective in both South Africa (ICER US$ 247/DALY averted, 95% UR US$ 135-815) and Uganda (US$ 96 per DALY averted, 95% UR US$ 54-265) compared to the standard of care based on SSM + existing diagnostics.

Shah 2013 also found AlereLAM + smear was highly cost-effective compared to the standard of care (SSM alone), with an ICER of US$ 29/DALY averted (95% UR $21-152). The more favorable ICER in the Shah 2013 analysis reflected both lower incremental costs for AlereLAM ($9 versus $21, largely reflecting assumptions about ALereLAM specificity) and higher incremental effectiveness (largely due to differences in study populations leading to different estimated life expectancy among HIV-positive TB survivors, 12.9 years versus 5.0 years).

When considering Xpert-based algorithms, Sun 2013 found that addition of Xpert + AlereLAM was highly cost-effective compared to Xpert + existing diagnostics, with an ICER in Uganda of $513/DALY averted (95% UR $164-8,707). Shah 2013 also found AlereLAM + Xpert highly cost-effective compared to Xpert alone, with an ICER of $45/DALY averted (95%UR $10-152). The reason for this more pronounced difference in cost-effectiveness is less clear but may reflect higher assumed case-finding in Sun 2013 reference scenario (93% [presumed] versus 85%, as above), leaving fewer individuals to be incrementally diagnosed by AlereLAM.

Sensitivity analyses

Four of the included studies presented sensitivity results. Across the four studies, models were robust to one-way sensitivity analysis across key parameters and ICERs were consistently cost-effective across most parameters investigated. Thus, the models generally agreed that influential parameters included: LF-LAM specificity, TB prevalence, and life expectancy after TB cure.

Orlando 2018 performed one-way and multi-way sensitivity analyses for the Xpert approach compared with the standard or care (SSM on symptom positives) and found prevalence of TB and cost of Xpert to be the most influential parameters, while ICERs remained consistently cost-effective across all parameters investigated. Sensitivity analysis for Xpert/AlereLAM approach was not reported.

Reddy 2019 performed one-way and multi-way deterministic sensitivity analyses, and found the intervention to be cost-effective (with less than 10% change in ICER) across nearly all parameters ranges explored in both countries; the one exception being the intervention diagnostic yield. Reddy 2019 found the intervention grew more cost-effective over time, and in sensitivity analysis cost-effectiveness at 2 years was not consistently below WTP thresholds. Interventions were cost-effective across nearly all values of empirical treatment probability and TB prevalence at 5 years, and when lifetime horizon was employed. In scenario analyses Reddy 2019 assess cost-effectiveness among patients with CD4 cell counts less than 100 cells/μL, while ICERs were higher across this group ($490 vs. $450 Malawi & $1,000 vs. $840 South Africa), the intervention remained cost-effective.

Sun 2013 found the parameters with greatest influence on the AlereLAM ICERs (in both South Africa and Uganda) were life expectancy after TB cure, the cost of TB treatment, LF-LAM specificity, and TB prevalence. One-way sensitivity analyses for Xpert-based scenarios were not presented in this study. Shah 2013 found the most influential parameters on AlereLAM ICERs relative to SSM were life expectancy after TB cure, sensitivity of clinical diagnosis, and TB prevalence. For AlereLAM relative to Xpert alone, the most influential parameters were the percentage of patients with low CD4 (<100), TB prevalence, and AlereLAM specificity. Sun 2013 performed three-way sensitivity analyses around these three parameters, finding that in both South Africa and Uganda, addition of AlereLAM to smear remained cost-effective when TB prevalence was as low as 5% assuming a life expectancy from 1.5 to 10 years and a specificity of 95% for AlereLAM.

In scenario analysis, Shah 2013 assessed the impact of inclusion of HIV/ART costs and effects. The ICER was higher (i.e., less favorable to AlereLAM) in all scenarios, but remained cost-effective $422 (95% UR $200- 752) for AlereLAM + SSM compared to SSM alone, and $367 (95% UR $163-646) for Xpert plus AlereLAM

compared to Xpert alone. Ultimately, the cost-effectiveness in this scenario converged to the cost-effectiveness of HIV care and ART. In the second scenario analysis Shah 2013 restricted the patient population to those with CD4<100 cells/μL. The ICER was even more cost-effective in this population due to low yield of SSM. For AlereLAM + SSM compared to SSM alone, the ICER was $25/ DALY averted, while the ICER comparing Xpert plus AlereLAM to Xpert alone fell to $35 /DALY averted.

Sun 2013 and Shah 2013 were the only two studies to perform probabilistic sensitivity analysis. Sun 2013 estimated that AlereLAM + smear + existing diagnostics would be cost-effective in >99.8% of simulations, at the WTP threshold of GDP per capita in both Uganda and South Africa. For AlereLAM + Xpert + existing diagnostics, Sun 2013 estimated 90% of simulations found AlereLAM to be cost-effective in South Africa at the GDP per capita WTP threshold, while an estimated 85% found AlereLAM to be cost-effective in Uganda. Shah 2013 presented cost-effectiveness acceptability estimates only for the scenarios with costs of HIV care and ART included. In these scenarios, Shah 2013 estimated AlereLAM + SSM, would be cost-effective in 72% of simulations, and AlereLAM+ Xpert, would be cost-effective in 77% of simulations. The lower probability of cost-effectiveness reflects the addition of costs for HIV care (including ART) that were not included by Sun 2013, who took a TB program perspective.

Principal Findings

- 6 eligible studies were identified all from sub-Saharan African settings with high TB/HIV burden.
- Models consistently demonstrated AlereLAM containing approaches could be cost- effective among African HIV positive adults across a range of settings and parameters evaluated despite heterogeneous diagnostic approaches evaluated.
- Mukora 2018 performed a detailed micro-costing of AlereLAM as part of point of care implementation and start-up in South African outpatient clinics, reporting unit test cost of $23.55 for AlereLAM several fold higher than previous estimates ranging from $3 to $3.99. The higher estimates are largely driven by the inclusion of above-clinic-level costs. If cost-effectiveness models are underestimating true costs of implementing AlereLAM, ICERs may be less favourable and programs more costly.
- Most models suggest cost-effectiveness may be improved among those with lower CD4 cells, Reddy 2019 found AlereLAM approach was actually less cost-effective among unselected hospitalized patients with lower CD4 cells counts, due to greater increased costs compared to all patients (costs associated with non- tuberculosis opportunistic diseases and concomitant increases in ART costs).
- With only a few studies and key differences in modeling approaches, assumptions, diagnostic algorithms assessed, analytical techniques, and study settings, applicability to other settings is limited.

Discussion

After performing a thorough review of the published and unpublished literature and broadening our inclusion criteria, we were able to identify only 6 eligible studies, all from sub-Saharan African countries with high TB/HIV burden. Studies were heterogenous in the range of diagnostic algorithms assessed and baseline

comparisons used, however models consistently demonstrated AlereLAM containing approaches were cost-effective across a range of settings and parameters evaluated.

Four studies calculated incremental cost-effectiveness ratios evaluating AlereLAM strategies. Two earlier studies were produced by the same research group using similar modelling approaches and assumptions and included in the previous systematic review by Hanrahan 2015[3]. Sun 2013 estimated ICERs of $247/DALY averted in South Africa and $96/DALY averted in Uganda using AlereLAM + SSM compared with SSM alone and $513 in South Africa and $231 in Uganda using AlereLAM + Xpert compared with Xpert alone. Sun 2013 targeted hospitalized patients with CD4 counts <100 cells/μL while Shah 2013 took a broader approach, testing inpatients and outpatients with presumptive TB irrespective of CD4 cell count. Shah 2013 estimated ICERs of $29/DALY averted in Uganda using AlereLAM + SSM compared with smear and $45/DALY averted using AlereLAM + Xpert compared with Xpert alone. Both studies performed scenario analyses to show that excluding costs of subsequent ART and HIV care showed AlereLAM approaches to be even more highly cost-effective. Both models identified AlereLAM specificity, life expectancy after TB cure and TB prevalence to be key influential variables on model results.

Two additional studies estimated ICERs and are included in this updated review. Orlando 2018 modelled cost-effectiveness among outpatients initiating ART in Mozambique, estimating an ICER of $72.31/DALY averted using an AlereLAM + Xpert diagnostic strategy stratified by CD4 cell count compared with SSM alone. While this algorithm is not directly comparable with previous approaches it does demonstrate that the addition of AlereLAM to diagnostic algorithms may be cost-effective across a variety of algorithms and implementation approaches, and was the first published study on AlereLAM cost-effectiveness from Mozambique. Orlando 2018 was also the first to account for additional costs of newly/additional transmitted infections due to delayed diagnosis and health system delay, although costs associated with HIV follow-up and care were not included. Inclusion of these costs resulted in higher total direct and indirect costs for standard of care approach compared with Xpert or AlereLAM + Xpert approaches investigated, therefore standard of care was dominated by intervention strategies (Xpert alone, then AlereLAM +Xpert). In this analysis Orlando 2018 found that the use of AlereLAM in their algorithm reduces the number of tests required only for those who are positive, therefore increasing the total number of tests required as AlereLAM negatives went on to subsequent Xpert testing.

Using data from the recently published STAMP trial[14] Reddy 2019 estimated among unselected hospitalized patients an ICER of $450/YLS in Malawi and $840/YLS in South Africa using Xpert + AlereLAM + urine Xpert compared with Xpert as the standard of care. Unlike previous models which generally used a SSM standard of care, Reddy 2019 compared AlereLAM interventions against a standard of care containing Xpert, therefore ICERs would be expected to be higher than when comparing against standard of care with lower diagnostic yield. In modified interventions of AlereLAM + Xpert, ICERs were at least as cost- effective and often improved. Reddy 2019 were able to calibrate models to STAMP trial data across the two countries to project impact on mortality and included two countries with vastly different economies and health system structures; low rates of sputum provision in Malawi coupled with high rates of empirical treatment in South Africa, meant the impact on diagnostic yield of AlereLAM was much greater in Malawi compared with South Africa. While previous work suggested cost-effectiveness may be improved among those with lower CD4 cells, Reddy 2019 found among hospitalized HIV patients, the AlereLAM approach was actually less cost-effective among those with lower CD4 cells counts, due to greater increased costs in this group compared to all patients (costs associated with non-tuberculosis opportunistic diseases and concomitant increases in ART costs).

Two additional studies performed costing analyses. Boyles 2018 used published unit test costs and the proportion of patients who would need each test using a given algorithm to calculate cost of investigation per patient screened. Costs ranged from $10.5 per patient with Xpert and culture algorithms to $42 per patient with AlereLAM/Xpert with sputum induction/culture approach. Boyles' 2018 population was restricted to hospitalized patients with cough and patients unable to produce sputum after induction were excluded,

possibly reflecting a sicker population that may have benefited from AlereLAM. Results may not be applicable to settings where sputum induction and culture are not available. Mukora 2018 performed a detailed micro-costing of AlereLAM as part of point of care implementation and start-up in South African outpatient clinics, reporting unit test cost of $23.55 for AlereLAM several fold higher than previous estimates ranging from $3 to $3.99. The higher estimates are largely driven by the inclusion of above-clinic-level costs required to support point of care implementation for AlereLAM, and staff costs were the primary driver of costs at both clinic and above clinic level. If cost-effectiveness models are underestimating true costs of implementing AlereLAM, ICERs may be less favourable and programs more costly. However, Mukora 2018 point out that as scale-up continues, economies of scale and other efficiencies may be gained.

Conclusions

Economic evidence for the implementation and scale-up of AlereLAM is still limited. Existing studies show a consistent trend, suggesting a high probability that AlereLAM could be cost-effective in a population of African adults living with HIV (particularly amongst hospitalized patients) but with only a few studies and key differences in modeling approaches, assumptions, diagnostic algorithms assessed, analytical techniques, and study settings, generalizability and more specifically, applicability to other settings is limited.

Inclusion of costs associated with ART and HIV care resulted in higher ICERs as TB diagnostic costs represented just a small proportion of total increased costs when HIV care is included. Models found cost-effectiveness of AlereLAM to be robust across a variety of sensitivity analyses, variations in key parameters and across different country settings and scenarios. Key parameters that are likely influential on cost-effectiveness include: TB prevalence, target population, and AlereLAM specificity, cost of treating TB and HIV and life expectancy post TB survival, and time horizon.

While current evidence is consistent in suggesting AlereLAM is likely cost-effective among HIV-positive patients in sub-Saharan Africa, caution should be used when extrapolating from a small number of studies, and additional evidence from a wider range of populations, settings and diagnostic approaches will be necessary.

Acknowledgements

I would like to acknowledge and thank Dr Sazia Naz and Olivia D'Silva for their time and efforts performing title and abstract review and for their conscientious assistance with data extraction.

Table 1: Characteristics of included studies

Study Characteristics	Sun 2013	Shah 2013	Boyles 2018	Mukora 2018	Orlando 2018	Reddy 2019
Country setting	South Africa & Uganda	Uganda	South Africa	South Africa	Mozambique	Malawi & South Africa
Year of cost valuation	2010	2013	Not Reported	2014	2016	2017
Currency	USD	USD	USD	USD	USD	USD
Clinical setting	Inpatient	Inpatient & Outpatient	Inpatient	Outpatient	Outpatient	Inpatient
Study population	Hospitalized HIV-positive adults with presumptive TB and CD4 count<100 cells/μL	HIV-positive adults with presumptive TB	Hospitalized HIV-positive, adults with presumptive TB	HIV-positive adults with CD4 ⩽150 cells/μL and who had not received ART or TB treatment in the preceding 6 or 3 months	HIV-positive adults initiating ART	Hospitalized HIV-positive adults irrespective of TB symptoms
AlereLAM diagnostic strategies	AlereLAM in addition to reference strategy	1) AlereLAM plus smear microscopy 2) AlereLAM plus Xpert	1) AlereLAM plus Xpert 2) AlereLAM plus Xpert and Culture 3) AlereLAM plus Xpert (induced sputum -SI) 4) AlereLAM plus Xpert SI and culture	AlereLAM & Haemoglobin (Hb) test	AlereLAM in all patients with CD4 cell count <200 & Xpert in all patients with CD4 cells count >200 and in those with CD4 cells count <200/ mm3 and negative AlereLAM results.	1) Xpert, AlereLAM and urine Xpert; 2) Xpert and AlereLAM
Reference diagnostic strategies	SSM or Xpert with clinical judgement and array of existing diagnostics	1)SSM 2)Xpert	1) Xpert and culture 2) Xpert SI and culture	N/A	1) SSM for those positive on 4 symptom screen 2) Xpert for all	Xpert
Analysis perspective	Public sector TB program	Health system	Health syste	Health system	Health syste	Health syste

Type of economic evaluation	Cost-utility	Cost-utility	Costing study	Costing study	Cost-utility	Cost effectiveness & Budget Impact Assessment
Source of costing	Non-empirical (utilizing existing costing sources)	Empirical (TB diagnostics), non-empirical (TB treatment)	Non-empirical (utilizing existing costing sources)	Empirical **(**Including clinic financial records, direct observation & completed timesheets)	Empirical costs (human resources associated with test delivery from program records) & Non-empirical (diagnostics costs from existing costing sources)	Mix of empirical and non-empirical (obtained data from STAMP trial-country-specific costing studies and national laboratory listings, diagnostics costs from existing costing sources))
Primary economic outcome	Incremental cost/DALY averted	Incremental cost/DALY averted	Cost/patient	Unit test cost	Incremental cost/DALY averted	Incremental cost/YLS
Type of model	Decision analysis	Decision analysis	N/A	N/A	Decision analysis	Modified CEPAC - I (microsimulation model)
Sensitivity analyses	One-way, multi-way and probabilistic	One-way and probabilistic	N/A	N/A	One-way sensitivity analysis	One-way and multi-way deterministic sensitivity analysis
Key scenarios/variables explored in sensitivity analyses	-AlereLAM specificity -TB prevalence -Life expectancy after TB cure Scenarios: -Xpert (instead of SSM)	-Inclusion of HIV-associated costs and effects -Restriction to CD4<100	N/A	N/A	- TB prevalence - Cost of Xpert test - Mortality among false negatives - Cost of symptom screen	- TB prevalence -% able to provide sputum -empiric treatment -diagnostic yield -cost of TB tests. Scenarios:

						- time horizons - alternative target populations (subgroup with CD4 cell count <100
WTP threshold	Per-capita South African 2012 GDP ($7275); Per-capita Ugandan 2012 GDP ($509)	Per-capita Ugandan 2013 GDP ($487)	N/A	N/A	Three times the per capita GDP Mozambique 2017 ($1,146 USD)	ICER of second line ART as determined using the CEPAC - I model: $750 USD /YLS in Malawi and $940 USD /YLS in South Africa

Abbreviations: USD, United States dollars; TB, tuberculosis; AlereLAM, lateral flow urine lipoarabinomannan test; Xpert, Xpert MTB/RIF; SOC: Standard of Care; DALY, disability adjusted life year; GDP, gross domestic product; WTP: Willing to pay; CEPAC-I;Cost-Effectiveness of Preventing AIDS Complications-International model; SSM, sputum smear microscopy; YLS, years of life saved

Table 2: Model parameters from included studies

	Value (sensitivity/uncertainty range)					
	Sun 2013	Shah 2013	Boyles 2018	Mukora 2018	Orlando 2018	Reddy 2019
Epidemiologic Parameters						
TB prevalence among symptomatic HIV patients (CD4<100 cells/μL)	0.38 (0.12-0.5)	0.3 (0.3-0.5)	0.51		0.101	0.235 Malawi 0.285 South Africa (0.10-.50)
TB prevalence among symptomatic HIV patients (CD4≥100 cells/μL)		0.1 (0.03-0.3)				
Prevalence of MDR-TB among new TB cases		0.014 (0.005-0.1)				0.01 Malawi 0.03 South Africa
Prevalence of MDR-TB among previously treated TB cases		0.12 (0.03-0.19)				
HIV patients with CD4< 200 cells/μL)					0.362	median CD4 count: 219 cell/ μL
Diagnostic Testing Parameters						
AlereLAM sensitivity	0.66 (0.3-1)	0.49 (0.39-0.59)	.355 (.28 - .43)		0.49	0.53 CD4 <100 0.42 CD4 ≥100
AlereLAM specificity	0.95 (0.7-1)	0.97 (0.9-1)	.933 (.88 - .96)		0.90	0.96 CD4 <100 0.98 CD4 ≥100
Smear microscopy sensitivity		.32 (0.3-0.51)			0.43	
Smear microscopy specificity		.99 (0.9-1)			1	
Sensitivity of clinical diagnosis		.3 (0-0.75)				
Specificity of clinical diagnosis		.89 (0.5-1)				

Xpert sensitivity	0.85	0.76 (0.41-1)	0.929 (Xpert Spot); 0.905 (Xpert SI)		0.976 (.93-.9976)	0.40 CD4 <100 0.43 CD4 ≥100
Xpert specificity	0.998 (0.848-1)	0.98 (0.93-1)	0.978 (Xpert Spot); 0.945 (Xpert SI)		0.992 (.984 - .9969)	0.99
Urine Xpert sensitivity						0.31 CD4 <100 0.13 CD4 ≥100
Urine Xpert specificity						0.99
Sensitivity of 4 symptom screen					0.775	
Specificity of 4 symptom screen					0.704	
Standard algorithm sensitivity	0.345 (0.2-0.5)					
Standard algorithm specificity	0.998 (0.848-1)					
AlereLAM cost	$3.78 ($1-32)	$3.64 ($2-10)	$3.50		$3.99 ($1.99 to $5.98)	$3 ($2-8)
Standard algorithm cost	$1.58 ($2-4)				$4.00 ($2 - $ 6)	
Smear cost		$0.87 ($1-$3)			$3.13 ($1.56-$4.69)	

Urine Xpert, concentrated cost						$26 – Malawi ($6-36); $15 - South Africa ($5-35)
Xpert MTB/RIF cost		$15.16 ($10-35)	$32		$14.72 ($7.36 - $22.08)	$25-Malawi; $15-South Africa ($5-35)
Treatment and Outcome Parameters						
Mortality of untreated smear positive TB	1	1 (0.75-1)				0.086 (monthly)
Mortality of untreated smear negative TB	1	1 (0.5-1)				
Mortality among false negatives (delayed diagnosis and treatment), %					0.2 (.10-.30)	
Mortality in those with TB treatment given (In care)	0.2 (0.1-0.3)	0.105 (0.04-0.3)			0.05 (0.025 – 0.075)	
TB treatment success rate (DS-TB)		0.77 (0.62-0.95)				0.95
TB treatment success rate (MDR-TB)						0.78
TB relapse (monthly range, based on time from treatment completion)						.00009-.0033
Treatment default (monthly, based on LTFU probabilities)						0.03-0.06

Life expectancy after TB recovery (years)	5 (1.5-10 yrs)	12.9 (1.5-33.5 yrs)			12.9 (6.450 - 19.350 yrs)	
Life expectancy with untreated TB with HIV (years)	0.0833 (0.071-0.25yrs)				0.665 (0.500 - 2.00)	
DS-TB treatment cost	$178 (Uganda) $850 South Africa ($500-2000)	$197 ($100-500)			$9.84	$42 (6 month regimen)
MDR-TB treatment cost, monthly (24-month duration)						$5544 (24 month regimen)
Cost of first-line ART (monthly, USD)						$11
Economic burden of a new TB case due delayed diagnosis and treatment					$847.00	
Disability weight, HIV on ART	0.167 (0.142-0.192)				0.053	
Disability weight, HIV not on ART	0.505 (0.085-0.115)					
Disability weight, TB with HIV infection		0.399 (0.267-0.547)			0.399	
Disability weight TB treatment	0.1 (0.085-0.115)	0.1 (0.028-0.115)			0.399	

Costs in USD: Mukora: 2014; Boyles: 2011 to 2013; Orlando: 2016; Reddy: 2017

Abbreviations: TB, tuberculosis; MDR-TB, multidrug resistant tuberculosis; DS-TB: Drug sensitive Tb AlereLAM, lateral flow lipoarabinomannan test; ART, antiretroviral treatment; HIV, HIV, human immunodeficiency virus; USD, United States Dollars. LTFU: Lost to follow up

Table 3. Cost and cost-effectiveness results (USD*)

Measure	Sun 2013	Shah 2013	Boyles 2018	Mukora 2018	Orlando 2018	Reddy 2019
Cost per patient (95% UR)						
Smear	$243 (South Africa) $71 (Uganda)	$62 ($37-116)				
Smear + AlereLAM	$308 (South Africa) $92 (Uganda)	$71 ($45-157)				
Xpert		$86 ($57-137)				
Xpert ->Culture			$10.5			
AlereLAM+Xpert		$91 ($60-163)				
AlereLAM ->Xpert			$10.5			
AlereLAM ->Xpert ->Culture			$12.5			
AlereLAM ->Xpert (SI)			$37.20			
Xpert (SI) -> Culture			$49.6			
ALereLAM ->Xpert SI ->Culture			$42			
AlereLAM + POC Hb				$24.93		
ALereLAM				$23.55 (US$11.49 clinic-level and US$12.06 above-clinic-level)		
ICER (Cost per DALY averted, 95%UR)						**(Cost/YLS)**
SSM	REF	REF			REF	
SSM + AlereLAM	$247 ($135-815) South Africa $96 ($54-265) Uganda	$29 ($21-152)				

Xpert	REF	REF			$56.54	REF
Xpert + AlereLAM	$513 ($164-8707) South Africa $231 ($125-3162) Uganda	$45 ($10-152)			$72.31	
Xpert + AlereLAM + Urire Xpert: all patients						$450: Malawi; $840: South Africa
Xpert + AlereLAM + Urire Xpert (patients with CD4<100/μL)						$490: Malawi; $1,000: South Africa
Sputum Xpert + Urine AlereLAM (all patients)						$420: Malawi; $810: South Africa

Abbreviations: UR, uncertainty range; AlereLAM, lateral flow urine lipoarabinomannan test; Xpert, Xpert MTB/RIF, DALY, disability adjusted life year; ICER, incremental cost-effectiveness ratio; YLS, year of life saved; ND, no data presented. NGO ¼ non-governmental organisation, NDoH ¼ National Department of Health

*Costs are reported in USD, year as per presented in analysis

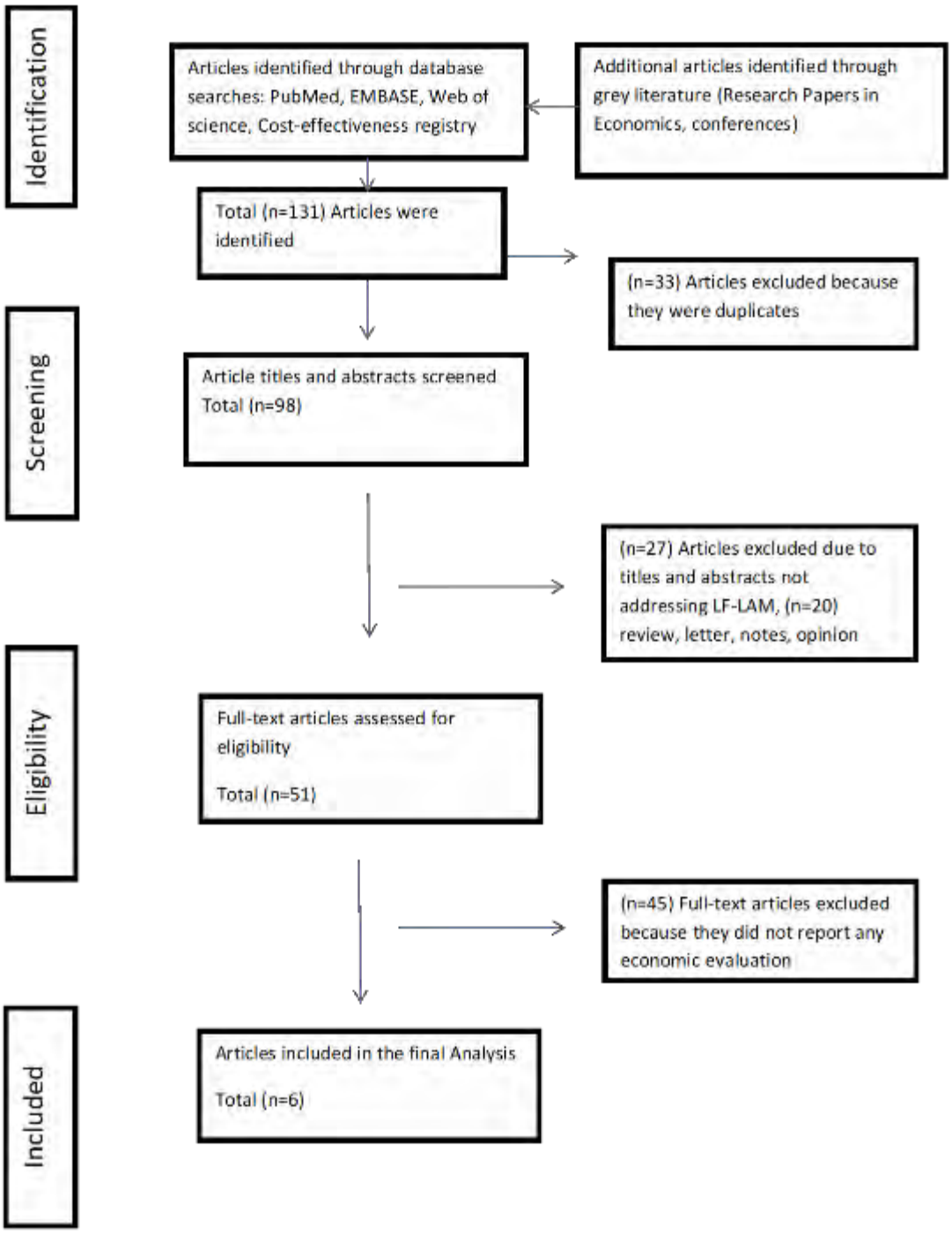


Figure:1 PRISMA flow diagram

Appendix Table 1: Quality of Health Economic Studies Instrument Assessment

Item	Sun 2013	Shah 2013	Boyles 2018	Mukora 2018	Orlando 2018	Reddy 2018
Is the study population clearly described?	Yes	Yes	Yes	Yes	Yes	Yes
Are competing alternatives clearly described?	Yes	Yes	Yes	Yes	Yes	Yes
Is a well-defined research question posed in answerable form?	Yes	Yes	Yes	Yes	Yes	Yes
Is the economic study design appropriate to the stated objective?	Yes	Yes	Yes	Yes	Yes	Yes
Is the chosen time horizon appropriate in order to include relevant costs and consequences?	Yes	Yes	N/A	N/A	Yes	Yes
Is the actual perspective chosen appropriate?	Yes	Yes	Yes	Yes	Yes	Yes
Are all important and relevant costs for each alternative identified?	Yes	Yes	Yes	Yes	Yes	Yes
Are all costs measured appropriately in physical units?	Yes	Yes	Yes	Yes	Yes	Yes
Are costs valued appropriately?	Yes	Yes	Yes	Yes	Yes	Yes
Are all important and relevant outcomes for each alternative identified?	Yes	Yes	Yes	Yes	Yes	Yes
Are all outcomes measured appropriately?	Yes	Yes	Yes	Yes	Yes	Yes
Are outcomes valued appropriately?	Yes	Yes	Yes	Yes	Yes	Yes
Is an incremental analysis of costs and outcomes of alternatives performed?	Yes	Yes	No	No	Yes	Yes
Are all future costs and outcomes discounted appropriately?	Yes	Yes	N/A	N/A	Yes	Yes

Are all important variables, whose values are uncertain, appropriately subjected to sensitivity analysis?	Yes	Yes	No	No	Yes	Yes
Do the conclusions follow from the data reported?	Yes	Yes	Yes	Yes	Yes	Yes
Does the study discuss the generalizability of the results to other settings and patient/client groups?	Yes	Yes	Yes	Yes	Yes	Yes
Does the article indicate that there is no potential conflict of interest of study researcher(s) and funder(s)?	Yes	Yes	Yes	Not mentioned	Not mentioned	Yes
Are ethical and distributional issues discussed appropriately?	Yes	Yes	Yes	Yes	Not mentioned	Not mentioned

References

1. WHO. 2018 Global TB Report.
2. WHO. The use of lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis and screening of active tuberculosis in people living with HIV: policy guidance. . Geneva: World Health Organization, 2015.
3. Hanrahan C, Dowdy D. A Systematic Review of Economic Evaluations of Lateral Flow Urine Lipoarabinomannan Assay for Diagnosis of Active Tuberculosis in HIV-infected Individuals: A Report for the World Health Organization, 2015.
4. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009; **6**(7): e1000100.
5. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; **6**(7): e1000097.
6. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ* 2013; **346**: f1049.
7. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value Health* 2013; **16**(2): 231-50.
8. Boyles TH, Griesel R, Stewart A, Mendelson M, Maartens G. Incremental yield and cost of urine Determine TB-LAM and sputum induction in seriously ill adults with HIV. *Int J Infect Dis* 2018; **75**: 67-73.
9. Mukora R, Tlali M, Monkwe S, et al. Cost of point-of-care lateral flow urine lipoarabinomannan antigen testing in HIV-positive adults in South Africa. *Int J Tuberc Lung Dis* 2018; **22**(9): 1082-7.
10. Orlando S, Triulzi I, Ciccacci F, et al. Delayed diagnosis and treatment of tuberculosis in HIV+ patients in Mozambique: A cost-effectiveness analysis of screening protocols based on four symptom screening, smear microscopy, urine LAM test and Xpert MTB/RIF. *PLoS One* 2018; **13**(7): e0200523.
11. Reddy KP, Gupta-Wright A, Fielding KL, et al. Cost-effectiveness of urine-based tuberculosis screening in hospitalised patients with HIV in Africa: a microsimulation modelling study. *Lancet Glob Health* 2019; **7**(2): e200-e8.
12. Shah M, Dowdy D, Joloba M, et al. Cost-effectiveness of novel algorithms for rapid diagnosis of tuberculosis in HIV-infected individuals in Uganda. *AIDS* 2013; **27**(18): 2883-92.
13. Sun D, Dorman S, Shah M, et al. Cost utility of lateral-flow urine lipoarabinomannan for tuberculosis diagnosis in HIV-infected African adults. *Int J Tuberc Lung Dis* 2013; **17**(4): 552-8.
14. Gupta-Wright A, Corbett EL, van Oosterhout JJ, et al. Rapid urine-based screening for tuberculosis in HIV-positive patients admitted to hospital in Africa (STAMP): a pragmatic, multicentre, parallel-group, double-blind, randomised controlled trial. *Lancet* 2018; **392**(10144): 292-301.

Web Annex D.14. User perspectives on LF-LAM for the diagnosis of active tuberculosis: results from qualitative research

ACKNOWLEDGEMENTS

Nora Engel and Muthoni Mwaura

Department of Health, Ethics & Society; Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, Netherlands (The Kingdom of)

n.engel@maastrichtuniversity.nl

Produced in preparation for the WHO guideline group meeting on TB-LAM in May 2019.

1. Introduction

In ensuring access to effective diagnostics for TB care, we not only need to assess that these technologies are accurate but also that they are feasible, useable and acceptable. The users of diagnostics include patients, clinic staff, lab managers, ministries of health, NGOs, regulators and suppliers. If we do not take the perspective of all users into consideration, we risk that these technologies do not fit their intended use setting, cannot be made to work and scaled up, are not utilized or not accessible for those in need. User perspectives on new diagnostics, their preferences and values as well as their experiences with existing diagnostic systems, are important to take into account during WHO decision-making on new diagnostics, including guideline development and policymaking. Feedback from representatives of key stakeholders groups (including patients, health professionals and programme managers) is important.

Studies generating this kind of data are often qualitative in nature (i.e. they focus on meanings that people bring to a phenomena and how they act upon it). Qualitative studies use targeted sampling methods to capture diagnostic experiences across a range of users, diseases, tests and diagnostic settings (Davids et al., 2015; Engel et al., 2017; Engel et al., 2015; Engel et al., 2018; A. McDowell & Pai, 2016; Andrew McDowell et al., 2018; Miller, Parkhurst, Peckham, & Singh, 2012; Squire et al., 2005; Yellappa et al., 2017). They are an ideal method for making sense of user experiences with and perspectives on diagnostic tools within “real-world” situations because they avoid placing assumptions about what these tools are expected to accomplish at the outset (e.g., that a test is easy to use). By involving users (e.g., through interviews, usability tests, ethnographies and user feedback), qualitative studies can support decision-making on diagnostics and offer concrete insights into users’ values and preferences, as well as acceptability and feasibility of new diagnostics in intended use setting. Such data will also point out important considerations for scale-up.

In May 2019, the World Health Organization will be evaluating two point-of-care tests for diagnosing TB in people with HIV (Abbott (formerly Alere)’s Determine TB LAM test and FujiLAM). To inform those discussions, the WHO has commissioned a study into the perspectives, preferences, and experiences of users of diagnostics (including people living with HIV and people affected by TB, health professionals,

and programme managers). To this end, we conducted a small qualitative study with participants in Kenya, Uganda, and South Africa. We interviewed clinicians, nurses, programme officers, laboratory staff, and patient advocates with the aim to understand their experiences of using TB LAM and diagnosing TB among people living with HIV (PLHIV) more generally and to contextualize users' preferences about a new diagnostic.

This study is exploratory in nature and part of an ongoing inquiry into user perspectives of new TB diagnostics. More, in-depth ethnographic research on the ground is warranted to better understand perspectives and practices of different users including PLHIV and their caregivers.

2. Methodology

In February and March 2019, NE and MW conducted 15 semi-structured interviews with clinicians, nurses, programme officers, laboratory staff, and patient advocates in Uganda, Kenya and South Africa. These countries were selected based on the fact they have policies in place regarding TB LAM. Due to the short timeframe participants were purposively sampled and approached based on convenience through personal contacts and colleagues. The majority of participants were from Uganda where TB LAM is already available in routine use (see table 1). It was not possible to speak directly to patients via the phone as most are seriously ill and even patient advocates did not know anybody who had tested with TB LAM. The advocates highlighted that the voice of seriously ill PLHIV are not well represented within the overall HIV advocacy. This warrants more in-depth and on the ground research with face to face interviews to understand all user perspectives and practices of diagnosing TB in PLHIV.

All but one of the interviews (which was done in person) were conducted via the phone. We asked for the testing and treatment experiences as well as experiences on interaction between providers and patients to contextualize users' preferences about a new diagnostic. Topics discussed included: current approach to diagnosing TB in PLHIV including specific challenges; experiences with using TB LAM, including details on steps taken in the diagnostic process, determining eligibility and treatment initiation as well as challenges and benefits; ways of interacting with patients about TB LAM; overall usefulness; the impact of TB LAM on equity and feasibility; and current policy context. We also tried to understand how a more complex test with longer turn-around time (TAT) (FujiLAM) would be perceived.

Interviews were audio-recorded, transcribed by MW, and coded by NE in NVivo. We each wrote memos on different topics, discussed these and collated them into themes which we present below. Professional roles are used to mask study participants' identity.

2. Ethics

This study was approved by UMREC, the ethical review board of Maastricht University. Study

participants were emailed an information sheet explaining the objectives of the study and an informed consent form which they signed prior to participation.

Table 1 Participants overview per country

	Uganda	Kenya	South Africa
Clinician	4	2	-
Nurse	1	-	1
Laboratory manager	1	1	-
Programme officer	1	1	-
Advocate	1	1	1

3. Results:

Below we discuss the results for current use of TB LAM separately for the three countries and then discuss overarching themes that emerged from the interviews across the different countries.

4. Current use of TB LAM in Uganda, Kenya and South Africa

Although TB LAM has been on the market since 2013, its adoption within national health systems of high TB/HIV burden countries has been very recent, if at all. As of 2018, South Africa, Kenya, and Uganda had each developed or began developing national guidelines for the test. South Africa's guidelines state that it should be used for all PLHIV in hospital settings and among those with CD4 counts less than 100/mm3 in primary care settings (TBCAB, 2018). Kenya's HIV programme recommends TB LAM be used as an adjunct rapid point-of-care diagnostic test for presumed TB among all PLHIV: (1) with advanced HIV disease (WHO stage 3 or 4 or CD4 count ≤ 200 cells/mm3 (or CD4% ≤ 25% for children ≤ 5 years)); or (2) any danger signs of severe illness; or (3) currently admitted to hospital (National AIDS and STI Control Programme, 2018). Uganda's TB guidelines recommend the use of TB LAM in HIV positive adults in whom TB has not been picked by microscopy or Xpert MTB/RIF and who are very ill, with a CD4 count of less than 100/mm3 (Uganda National Tuberculosis and Leprosy Control Program, 2017).

National roll-out of TB LAM varies between the three countries. Uganda is the only one thus far to have rolled out the test to national and regional-level referral hospitals, but according to participants of the study, not all districts have received the test, and stock-outs in those that have it have been experienced. Actual in-country usage also seems to vary. According to a lab manager at a teaching hospital, TB LAM is being used for both HIV and non-HIV immunosuppressed patients who are suspected of having TB and are not able to expectorate sputum while a clinician working at a national referral hospital only uses TB LAM when other tests are not able to detect but the clinical suspicion is still high. Where the test is conducted also varies as some settings prefer the test be done in the lab due to frequent change-overs of ward staff, while others prefer to do it by the bedside. Although the Ugandan guidelines recommend a CD4 cut-off of 100/mm3, some hospitals appear to be using a cut-off of 200/mm3 (ID6, nurse 1) or conducting the test irrespective of CD4 counts (ID8, clinician 4). This may reflect the leeway the TB programme seems to have given local settings to run the system in the way that suits the context best (ID1, programme officer 1).

The Kenya TB programme will begin to roll out the test to county-level referral hospitals as a pilot project in 12 high-burden TB/HIV counties. The algorithm will recommend the test be used in conjunction with Xpert among PLHIV in hospital settings, with CD4 counts of less than 200/mm3. This criterion was extended from WHO's cut-off of 100/mm3 on the rationale that expanding it will capture more patients and that in Kenya there is generally a good rate of adherence to ARV medicine, so limiting CD4s to WHO's recommendation will only capture a handful of patients (ID15, programme officer 2). According to a TB programme manager, TB LAM will be conducted in the lab to enable uniformity in result interpretation, be close to GeneXpert machines, and streamline recording and surveillance practices.

According to a presentation given by Dr. Lindiwe Mvusi from the Ministry of Health during the TB 2018 pre-conference held on Sunday 22 July 2018 (Mvusi, 2018), South Africa has developed an algorithm for TB LAM and has rolled out the test as a pilot project in five hospitals to be used concurrently with Xpert MTB/RIF. According to an advocate of the present study, data from the pilot project are still in review.

5. TB LAM makes a difference in a hard to diagnose patient group

Participants discussed the difficulties in diagnosing TB in PLHIV in their settings, which is often extra- pulmonary TB, and how the introduction of TB LAM has improved on this. An advocate working in a high burden TB/HIV district in Uganda, for instance, states the critical difference the test makes in a hard to diagnose patient group with high numbers:

> *"it's still an important test because we see like for the HIV in Uganda, we have about 89% of people on treatment but still we see like 10% of those having advanced HIV. And it is still the leading cause of mortality and morbidity among PLHIV, so definitely we still need it because with the previous technologies we had we are not diagnosing enough, we still have capacity issues, and I feel we need it, though it has to be used in combination with other technologies, and it can't be used in other populations but these populations are critical, and the numbers are high" (ID13, advocate 3).*

Although TB case-finding among this population may have improved, follow-up testing is still difficult and clinical observation is particularly challenging as this population is vulnerable to co-morbidities and drug interactions (ID8 clinician 4). Furthermore, the characteristics of this patient group, namely being very ill, means that typically the TB LAM test is not explained and consent not taken. Only if the result is found positive a clinician might then say we tested (without going into details of the test) and are pretty sure you have TB. If the results are negative, clinicians would not mention it (ID2 clinician 1, ID8 clinician 4). Since patients are very ill and admitted in the hospital, clinicians work with implied consent and there is time to discuss some of the common patients concerns about their diagnosis and what the implications are for treatment, side effect, pill burden and transmission to others (ID10, clinician 5). In a regional referral hospital in Uganda, these concerns are then discussed with the nurses during counseling for TB treatment (ID6 nurse 1).

6. Characteristics of the test

S a m p l e

Diagnosing TB in PLHIV is challenging as obtaining a viable sputum sample is often difficult because the patient is too ill to cough, or the disease is disseminated and the sputum sample may test negative. For these reasons, most of our study participants acknowledged the benefits of using urine to test for TB, citing it as a safe, pain-free, and non-invasive method for testing for TB that is easier to obtain than sputum. A nurse from Uganda illustrates these advantages on mortality (ID6, nurse 1):

> *"the challenges were, of course we were missing many cases, mostly these people who have HIV, they come in in their 3rd/4th stage, they cannot cough, they are not able to do the chest X-ray, they don't have strength to stand to take them for the chest X-ray, and you just treat blindly. Most of the people died because, we didn't know the diagnosis… because you can't give HIV cases, who are really sick, bed-ridden, and the cough is mild, it is not strong for you to conclude that this patient has TB. [But] the LAM has impacted that such that these bed-ridden ones, we are having early diagnosis, and early treatment, with less mortality now in HIV/TB" (ID6, nurse 1)*

A lab manager notes that obtaining urine instead of sputum from very ill patients does increase patient participation, as most are able to produce the latter over the former (ID9, lab manager 1). Additionally, a clinician emphasizes that when compared to sputum, urine presents less of an occupational health hazard to health workers and is less stigmatizing for patients (ID2, clinician 1). That being said, obtaining urine was not always easy. A clinician and a nurse noted that, at times, obtaining urine is a challenge when the patient is too ill or septic to produce it, when he/she has to be catheterized because collecting urine from diapers is impossible (ID7, clinician 3; ID12, nurse 2), or if the patient is in a hospital where there is no private and clean space to produce urine which is

common in rural hospitals in Uganda (ID7, clinician 3). A lab manager highlights how he is not always sure how old the sample is and whether what he receives in the lab is a fresh urine sample (ID14, lab manager 2). The non-invasive nature of the sample also allows testing without explicit consent from the patient (see above).

Turn- around time

The fast turnaround time (TAT) of TB LAM was often cited by participants to have a notable impact within their settings. With a running time of 25 minutes, it was frequently discussed how treatment can be initiated sooner than if a test was run using existing technologies (ID1; ID12; ID13; ID14; ID6; ID7; ID9; ID4). This in turn was linked to reduced loss to follow-up as patients do not have to wait extended periods of time for a diagnosis and treatment initiation. As a programme officer illustrates,

> *"If I am very ill, if my clinician can get a result in around 30 minutes, it basically makes the whole difference…I am in the ward very sick, I need to know [my] condition and then start treatment" (ID1, programme officer 1).*

A lab manager also notes that due to the workload within the labs, clinicians in his setting prefer to request for TB LAM over smear microscopy (ID14, lab manager 2).

However, while the running time of TB LAM is standard, the time it takes between collecting the urine sample for testing and initiating treatment varies based on the reporting system, availability of anti-TB drugs in the pharmacy, and the time of day the test is requested and conducted. Once the decision is made to initiate treatment, it can be commenced within a few hours if the drugs are available at the pharmacy (ID12, nurse 2; ID1, programme officer 1). Yet, in several hospitals it seems to be the next day if clinicians have already finished their ward rounds for the day. So even if the LAM test is done near bedside, treatment might take another day to be initiated. It was also mentioned that if the test is done in the lab, the TAT would be faster if someone follows-up directly with the lab versus if they wait for normal reporting processes (ID10, clinician 5). A lab manager in Uganda noted that if the report is ready after the clinician has completed ward rounds -which in this particular hospital end at 1 or 2pm-, diagnosis and treatment initiation can only begin the next day (ID14, lab manager 2).

User-friedliness

TB LAM was frequently referenced as straightforward and easy-to-use, often likened to using a pregnancy dipstick test (ID2, clinician 1; ID7, clinician 3; ID11, clinician 6). The lack of technical expertise required to run the test was said to allow for task sharing especially in settings where the workload of the laboratory technicians is very high (ID1, programme officer 1; ID5, advocate 2; ID9, lab manager 1). Some found interpreting results to be straightforward, and appreciated the graded scorecard that accompanied the test kit (ID14, lab manager 2).

However, each of these benefits was not without challenges. For example, while the test is not technical, a lab manager mentioned that its timing is vital and that going beyond the recommended time could affect the results (ID14, lab manager 2). TAT was also said to influence the number of tests that could be run concurrently (ID12, nurse 2) and having a timer on while running the test was important (ID5 advocate 2; ID12, nurse 2; ID7, clinician 3). Additionally, the simplicity and specific timing of the test was mentioned to influence who could run the test, as those with a lot to do in their

daily routine -such as clinicians- may forget they have began running a test and leave it to run longer than recommended (ID12, nurse 2).

Not everybody thought that interpretation of results which depends on visibility of the graded bands was easy. Several mentioned challenges with reading faint results, especially grade 1 and deciding on the result (ID2, clinician 1; ID8, clinician 4; ID9, lab manager 1; ID14, lab manager 2). In some settings this influenced whether the test was to be conducted in the laboratory or at the bedside, as the former was deemed a better environment for maintaining uniformity in result interpretation (ID2, clinician 1; ID15, programme officer 2), potentially affecting the future point-of-care status of the test.

Similarly, while the simplicity of TB LAM may enable non-laboratory staff to conduct the test, task sharing may not be feasible in settings where frontline staff conducting the test are frequently rotated through the system (e.g. nurses, students, clinical officers) (ID6, nurse 1). This could affect who in the end is able to conduct the test and how close to the bedside/patient it will be.

Lastly, although volume control for the test was perceived by some to be straightforward (ID14, lab manager 2), the fact that the kit does not come with a micropipette presents a challenge for others (ID8, clinician 4). It was therefore suggested that the manufacturer could provide detailed instructions of how to measure urine if a micropipette is unavailable (e.g. number of drops).*Cost and maintenance*

When it comes to the logistics surrounding TB LAM, it was perceived by most participants to be better suited for their settings than existing technologies. For example, unlike Xpert MTB/RIF, TB LAM: (1) is cheap to buy and maintain (ID3); (2) does not require other reagents (ID13); (3) does not require much lab space (ID11); (4) does not require cold chain (ID14), and (5) does not require electricity to run (ID5). For these reasons, many understood the test to address infrastructural and logistical issues that currently prevent other technologies from being used at optimal capacity (ID7, 13, 14).

That being said, the shelf-life of the test was perceived to be relatively short by some (ID1, 11), with a nurse recounting a recent instance when an expired test was used and the result was negative but a Xpert MTB/RIF result was positive (ID6, nurse 1). Additionally, accessories for running the test that do not come with the kit such as urine containers and micropipettes, presented challenges when not locally available (ID14, lab manager 2) and may lead to improvisation that could impact the reliability of results. Lastly, though TB LAM was largely perceived as inexpensive to procure, various settings in Uganda have experience stock-outs (ID6, nurse 1; ID14, lab manager 2), while private practitioners using the test in Kenya believe the test to still be too expensive for their patients (ID4, clinician 2; ID11, clinician 6).

7. Constructing confidence in diagnosing TB in PLHIV

Limited confidence due to sensitivity, cross reactions and faint results

The sensitivity of TB LAM limits the confidence in its results (ID3; ID5; ID11; ID13; ID2; ID4). According to a programme officer in Uganda, the confidence of clinicians in TB LAM is rather low given its low sensitivity that only improves with low CD4 count. This confidence decreases further because TB LAM can also give positive results due to cross-contamination in patients with candidiasis or with nontuberculous mycobacteria (ID1 programme officer 1). Furthermore, if somebody is weakly positive on TB LAM, when do you establish somebody as positive? Having the confidence that a grade 1 result is indeed a positive result is not given and according to a Ugandan programme officer some people argue that grade 1 should not be treated as a positive result (ID1, programme officer 1). While some mentioned they did not have any doubts in reading the results

(ID12, nurse 2), others ensure coherence and consistency in reading results and interpretation of grade 1 by conducting the LAM in the laboratory by lab technicians who have established a routine as opposed to rotating clinicians (ID2, clinician 1). Yet, there is no additional microbiological confirmation of the TB diagnosis, only in a few cases can Xpert Ultra be done (i.e. patients are able to produce a sputum) (ID2, clinician 1).

Test is not made to stand alone

In communication about TB LAM with her patients, a nurse in South Africa explains how she is managing expectations of results among patients. She always establishes first whether a person had TB before or has been tested and then explains that her test only uses urine. Patients generally want to know what is going on with them and view TB LAM as one more step towards finding out and getting better. Upon receiving a positive TB diagnosis (on top of being HIV positive), some are disappointed and sad about the double diagnosis (which is particularly tricky in settings with double stigma of HIV/TB (Daftary, 2012)), others are accepting knowing they are very ill and these things are possible (ID12, nurse 2). According to an advocate, patients have the tendency to believe test results over those from clinical diagnosis (ID13, advocate 3). The nurse in a South African hospital explains the uncertainty of the LAM results and refers to the doctor who will come later on during the day and make a decision, or order further tests. In doing so, the test is not made to stand by itself but embedded in a battery of tests and considerations that the doctor makes, and not the nurse who is conducting the test (ID12, nurse 2).

> *"I do explain to the patient that even if my test is negative, the doctors will wait for the other tests that they have done because my test doesn't mean that there isn't TB in the body, it just means that the test that I am doing cannot pick it up. Not to say that they do have TB but the possibility still does exist because I need to make ease as well that I am not coming there and saying they do have TB. The possibility is there and that is why they want to test them." (ID12, nurse 2)*

Interpreting the result with confidence: clinical suspicion is trump

Most clinicians we talked to seem to only act on a positive TB LAM result if they already suspect TB due to clinical presentation or symptom screen.

> *"if I am struggling to confirm a diagnosis of TB, and the LAM comes back positive and my clinical history fits, that may mean the life or death of that patient. Because that would mean I start my treatment sooner rather than later." (ID4, clinician 2)*

In Uganda, the TB programme directs clinicians that just having a low CD4 count is not enough for acting on a positive LAM result, they also need to have done the symptom screen and need to be suspecting TB so that they interpret a positive TB LAM result with confidence (ID1, programme officer 1). Backing up a positive LAM result with other follow up tests is hardly ever done, especially because the patients were not able to provide that sample (and that might have been the reason for doing LAM in the first place).

> "*Usually what happens is that these clinicians are telling us that by the time they are asking for a lab, they have already presumed TB. So even in the even in the event of cross contaminations, they are ready to believe that this is TB because the person is already presenting with the TB symptoms. That is why we are telling our clinicians that a low CD4 or me being very ill is not the air ticket to a TB LAM, no, I should be a presumptive TB case. I might have a very low CD4 but you need to screen me for TB, and if I have the signs and symptoms, then you go ahead and do the LAM*." (ID1, programme officer 1)

A clinician in a Kenyan private hospital would still try to confirm a positive TB lam result with other ways of looking for TB

> "*You know TB is just so difficult to diagnose, it [TB LAM] is an additional tool to our honorarium of TB diagnostics. Right, so it's just that. And it's actually a really good one because if it's negative, then it's less likely to be TB and that really helps I think*" (ID11, clinician 6).

A nurse in research study in South Africa and a clinician involved in a study in a district hospital Uganda echoed that sentiment; both observed that doctors would not start everybody on treatment with a positive TB LAM result, but wait for other evidence (a sputum sample, a culture result) if the patient is asymptomatic (ID12, nurse 2, ID2, clinician 1). A clinician argues that empirical suspicion will trump also a negative TB LAM result (ID2, clinician 1).

Treating severely ill patients with improved level of confidence

TB LAM results are particularly reassuring (to be sure it is indeed TB) in patients that are already very ill, have several co-morbidities and are therefore also much more susceptible to side effects and severe complications during TB treatment (ID2, clinician 1, ID8, clinician 4). That particular usefulness of TB LAM over empirical treatment might change if TB LAM would be made available for patients that are not so ill or do not have as many co-morbidities. In those patients, a clinician suspects that one would feel more comfortable to treat empirically and monitor whether indeed the patient improves (ID2, clinician 1).

> *"., around the world lots of patients are treated empirically for TB and it's kind of like, you are not very sick but you know to be honest 6 months of these drugs are well-tolerated, here you go, and we will see, you are going to go through 6 months of treatment and that's ok. That happens even in NYC, where you know, we are 70% sure you have TB but not 100% sure, and you know what its ok. I mean you will take these drugs for 6 months, and we will check your liver function a couple times and you will call us if you have any symptoms but you will be ok. And that's just not the case when someone is sort of like super sick. That's exactly who get all the side effects of the drug and will have a good chance like you say, multiple comorbidities going on, even if they truly do have TB they may very well have something else as well." (ID2, clinician 1)*

Global guideline makers' implicit confidence influences national and local level confidence in diagnostic

According to an advocate, the language around innovations such as new diagnostics and drugs used in guidelines and communication from the WHO heavily influences country uptake, particularly in countries without progressive HIV or TB programme managers. A wait and see attitude is then taking place. The advocate particularly laments the fact that operational research on TB LAM and its effects, which could have changed that attitude, was ignored for several years (ID5 advocate 2).

> "*the people of WHO, their attitude towards TB LAM Alere, was not very positive, (..) and they partly contributed to low up take of the LAM, because they overemphasized talking about the lack of performance, and didn't refer to the mortality benefit that we saw from especially the South African data. (…) if you as someone in the WHO has developed these guidelines, but you are not very convinced about the test… See now they are very excited about FujiLAM but when it was TB LAM they were like, "hmmm". It trickles on to the countries, it really does." (…)* "*I don't necessarily think it is bad that they applied the caution that they applied, but it was very wrong of them to ignore operational research for, how many years, 4 years, waiting for FujiLAM. This is nonsense. If operational research is out there, it is showing that there is a mortality benefit, it is showing you that there is use of the test for those with CD4 counts of less than 200 especially, including in ambulatory*

> *settings, its showing you that there is task sharing that can be done, that is on WHO's head…it was wrong of them to do that. Actually, ethically, it was very wrong, you see. That also slowed down the uptake of TB LAM Alere. Why should people die because people are waiting for a more specific and sensitive test?*" (ID5 advocate 2)

8. Eligibility criteria and CD4

A clinician and researcher of LAM argues that everyone with HIV in hospital settings should get a TB LAM, whereas the usefulness for outpatient settings is not clear yet. She guesses that everybody who looks sick in outpatient should get a LAM as well.

> *"..TB is often missed particularly in inpatient settings. So, I think inpatients settings, anyone who is HIV positive probably deserves a LAM test to make sure they don't have disseminated disease that could be better diagnosed with LAM. In outpatient settings, how you use LAM I think isn't well understood." (…) "I think that if you have someone presenting to clinics who are HIV positive, who look sick, you should certainly do a LAM test, if you are considering hospitalizing them. So, the people who look relatively well who are positive, you can do some kind of a test to understand what their baseline CD4 is*" (ID7, clinician 3).

It would need more in-depth research on the ground to understand how different healthcare providers decide when somebody is serious ill.

CD4 counts are not routinely available

According to a clinician in Uganda, LAM has been made conditional on another test that is not routinely available (ID7, clinician 3). CD4 counts are not routinely done in Uganda (ID1, programme officer 1) and it is very rare a clinician would base their LAM request on a CD4 count (ID1, programme officer 1; ID14, lab manager 2), even if available in a hospital laboratory, in order to avoid delays of a couple days and because patients generally look sick (ID14, lab manager 2). In a district hospital everybody who is admitted and HIV positive is tested on LAM irrespective of CD4 count (ID8, clinician 4). In a private hospital in Kenya where CD4 counting is easily available, the clinician would order the CD4 count and TB LAM concurrently but based on clinical suspicion. This is done to avoid delays when ordering a CD4 count which takes 1 or 2 days. Also, being a TB endemic country, the clinician has seen TB with all sorts of CD4 counts (ID11, clinician 6).

Unintended effects of CD4 below 100 cut-off

According to an advocate, the restrictive use of CD4 count below 100 had the undesired effect that countries would not like to admit that they have many patients that are that ill, because it reveals poor HIV programme performance (ID5, advocate 2):

> *"..about 30% of PLHIV in high HIV endemic countries, about 30% will have advanced HIV, in a lot of these low- and middle-income countries. So, they don't want to roll out a test that will show the enormity of the problem, ..(ID5, advocate 2).*

It also gave the national programme officers the opportunity to argue that these are very few people and therefore downplay the priority of the test which is why, according to an advocate, it is important to follow operational research results and widen eligibility criteria to above CD4 count of 200 (ID5, advocate 2). Given the restrictive eligibility and niche applicability of the test, it might be further downgraded in programming and budgeting for it (ID3, advocate 1) and not be made as accessible as it should be, according to an advocate in Uganda only 25% of those eligible receive a TB Lam, and this might have to do with the niche and strict eligibility criteria. It also lowered confidence among clinicians who initially are unsure about who the test is for and what results mean (ID14, lab

manager 2). An advocate would like to see LAM used in smaller city-level hospitals as well, not just the big hospitals (ID3, advocate 1). But seriously ill PLHIV that are mostly affected by LAM are not the ones that usually are represented well by advocacy of PLHIV:

> "*..and unfortunately, with LAM it is not one of those that can you get patients groups to shout*
> *too much about, because what do you say. Like we've said yes, we want LAM now blah blah, but PLHIV, those people who know what they want because they know that the treatment works, in those groups, they don't always deal with the advanced cases.* " (ID3, advocate 1)

9. Fuji LAM

Weighing TAT, complexity and sensitivity

A Kenyan programme officer argues that finding a right balance between the existing and future TB LAM test involves finding a balance between the applicability of the test in terms of TAT and accuracy (like between Xpert MTB/Rif and Ultra) and also depending on where a country is in its decision-making with adopting TB LAM. A less sensitive test may still be useful if it is operational within the country. "Sensitivity depends on what a country wants" is how he put it (ID15, programme officer 2).

Clinicians and lab managers seem to value higher sensitivity, especially in settings where the Alere/Abott TB LAM is presently conducted in the laboratory and not at the bedside (ID11, clinician 6; ID14 lab manager 2)). A Ugandan programme officer argues that additional user steps (which he envisions to be still easier than doing a microscopy) are a small price to pay for a higher sensitivity, even if this means it defeats the purpose of POC and needs to be moved back to the lab (ID1, programme officer 1). Others specify and argue a longer waiting time up to 2 hours would be acceptable but many more complex user steps (including for instance amplification) and moving it to a central lab would mean the strength of the test would be lost (ID2, clinician 1). A clinician argues she is more interested in a sensitive test if patients are in a hospital setting with a very short TAT to initiate treatment with the patient still in front of you (ID10; clinician 5). Additional complexity might also make the test not feasible for small labs attached to wards in terms of the required equipment, quality controls, staff capacity, sample preparation steps and overall cost (ID4, clinician 2).

An advocate argues that a better performing test will generally be picked up quicker by policymakers (ID5, advocate 2). Another advocate cautions that a more complex and longer TAT would mean programmatic drawbacks, as these aspects have been essential advantages of the ALere/Abott TB LAM (ID13, advocate 3). If the test should be rolled out in primary care settings a higher complexity is not warranted (ID 5, advocate 2).

TAT is also linked to how many tests a healthcare provider can run simultaneously and when they are run during a working day. A nurse in South Africa explains that if TAT would increase to an hour she would be able to run more than two tests simultaneously (currently she conducts maximum two at the same time). Yet, additional user steps might eat into that time again. (ID12, nurse 2). A clinician refers to the difference between ZN and LED microscopy where the additional handling and user steps mean the testing is often done in batches at the end of the working day to run them simultaneously, increasing overall TAT to a day (ID7, clinician 3).

10. Possible reasons for slow policy implementation and how to overcome it

Although TB LAM has been on the market for about six years, the uptake of the test within national, high-burdened health systems has been remarkably slow. When asked why they think this

has been the case, many participants cited the perceived low performance of the test by clinicians and policymakers (ID4, clinician2; ID5, advocate 2), largely shaped by the initial communication about the test published by WHO and related agencies (ID5 advocate 2). Even once guidelines for TB LAM were published, prioritizing the operationalization of those guidelines did not happen, unless there was high-level advocacy taking place as well (ID3 advocate 1, 5 advocate 2, 13 advocate 3). In many settings, this type of advocacy seems to be missing due to the absence of strong advocacy voices for this particular patient group (TB among *advanced* HIV individuals; ID5 advocate 2) or the lack of awareness among frontline healthcare workers that such a test even exists (ID4 clinician 2, 13 advocate 3). Additionally, the cumbersome process of operationalizing global guidelines and developing context-specific algorithms may discourage many resource-tight TB/HIV programs (ID7 clinician 3, ID15 programme officer 2). For this reason, countries may be awaiting operational research from those settings that are currently implementing it before adopting it within their own systems (ID3 advocate 1, 9 lab manager 1). It was also speculated that perhaps national programs have not prioritized TB case-finding among severely ill PLHIV. As a programme officer put it:

> "*if you really think that this [TB] is really a disease of serious public health importance amongst the PLHIVs, you would use all necessary tests to make sure that you identify this in this group*" (ID1 programme officer 1).

To overcome the relatively slow uptake of life-saving technologies, they must first be prioritized by policy-makers and the language within the policy should be carefully constructed so as to not unduly or unintentionally discourage uptake. At a global level, there should be greater integration between relevant programs surrounding the communication of such a test. For example, although TB LAM benefits both TB and HIV populations, if only the global TB or the global HIV programme provides communication regarding the test, the other programme at the national level will assume that it is not its responsibility. Both global HIV and TB programmes should communicate jointly and in a coordinated fashion as to clearly indicate responsibilities. Lastly, implementing partners could also partner with national TB/HIV programs to sponsor the operationalization of international guidelines and the development of accompanying reporting and surveillance tools.

11. Conclusion and recommendations user perspective TB LAM

The results show that TB LAM clearly addresses a need and makes a difference in a population in which TB is very hard to diagnose. The characteristics of the test, the sample, TAT, ease of use, cost and infrastructure/maintenance requirements are differently discussed among the participants. While global health actors including the participants of this study generally herald TB LAM as an easy to use, low maintenance/equipment requiring, quick test that crucially does not rely on sputum but a much more easily available and safer sample (urine), those very same characteristics can also pose their specific challenges as experiences of those using the test show. The sample, for instance, is safe, more easily available, and less stigmatized than sputum, but not everybody can collect it (in bedridden patients catheters are required); produce it (dehydrated patients with sepsis cannot urinate); adequately measure it (in some instances the dropper provided in the test kit was not accurate enough and a micro- pipette was required) or has a private and clean space to provide it (rural hospitals do not necessarily have toilets or running water available to patients). Similarly, the infrastructure requirements are minimal but stock outs, lack of urine containers and shelf live still pose challenges. While the TAT is supposed to be just 25 minutes, treatment initiation in many settings only happens the next day. Another important challenge that users struggle with is

the low sensitivity of the test, cross-reactions and the difficulty of reading faint results (grade 1). We show how clinicians and nurses construct confidence in the results by 1) ensuring that the test is not made to stand by itself but embedded in a battery of tests and considerations that the doctor makes, and not the nurse who is conducting the test; and 2) using test results in combination with clinical suspicion of TB or other evidence in case of asymptomatic patients. Could the implications of these practices mean that patients are still being missed?

And yet, in severely ill patients where pill burden, side effects and severe complications during treatment are a real challenge, TB LAM provides some much-needed confidence beyond clinical suspicion. Would this need for increased level of confidence change if the test would be made available to less ill patients in outpatient settings? Answering these questions would require more in-depth and on the ground research.

In the way confidence in the test is perceived, the global guideline making by WHO and the language that is used has very important consequences that trickle down to country and user levels. Several participants blame the negative language for the slow uptake of TB LAM and general hesitation by countries to implement. In the future, it should be carefully considered how guidelines around a new test are being communicated and drawbacks and benefits weighed against each other and presented. Both global HIV and TB programmes should communicate jointly and clearly indicate responsibilities. Prior consultations with users and policymakers could aide that process.

Our results also reveal that the restrictive eligibility criteria of the WHO guidelines mean the test is perceived as so niche and at the same time carries the implicit risk of revealing poor performance of HIV programs that it is not made accessible to the extent it should be. What is more, CD4 counting is not widely used to determine eligibility for LAM testing. Currently symptom screen, hospitalization and assessing whether a patient looks ill are used to decide whether a patient is eligible for TB LAM testing rather than CD4 count which in many places is not routinely available or if available not ordered to avoid time delays.

While our participants would value improved sensitivity in a new test such as Fuji LAM, they also caution against increasing complexity and TAT if the test should be made to work in primary care settings.

12. Recommendations to include qualitative research into guideline making on diagnostics

The GRADE-CERQual approach provides guidance for assessing synthesized qualitative evidence. The WHO has formally commissioned and included qualitative evidence into the Optimize guidelines on healthworkers role for maternal and newborn health, including thematic analysis of an email discussion list, in-depth case studies of country programs and four systematic reviews of qualitative evidence (Colvin, 2014; World Health Oganization, 2012). Since then, similar qualitative evidence has been used for several other guidelines by WHO (f.i. on healthworkers role in providing safe abortions; use of ARVs for treating and preventing HIV infection; antenatal care guidelines; health promotion interventions for maternal and newborn health, etc.). Note, that at times the qualitative evidence synthesized is from similar interventions rather than the exact same intervention (f.i. experiences with task shifting or adherence to treatment in related fields). This is useful to keep in mind for decision-making on new diagnostics, where qualitative studies on the utilization of the specific diagnostic in question are scarce and thus insufficient qualitative evidence is available to synthesize.

To overcome these limitations and to generate evidence on end-user and professional user experiences, preferences and values, three measures are proposed:

1. **Identify and acknowledge all cadres of users**: Focus on user perspectives and experiences that include end-users (such as patients), but also professional users such as laboratory technicians, clinicians, nurses, local suppliers and decision-makers (Shah, Robinson, & AlShawi, 2009).
2. **Engage users in decision-making about diagnostics**: Commission qualitative studies that draw

   on their perspectives and experiences to support WHO decision-making process around new diagnostic guidelines: assign a technical team to prepare a file on user experience, values and preferences using: qualitative evidence synthesis (where evidence is available); interviews/FGDs/ethnographies with user groups; in-depth case studies of country programmes, trials, or demonstration studies; moderate and conduct thematic analysis of online discussions forums;
3. **Mandate qualitative evaluations for diagnostic regulatory approval:** Mandate qualitative evidence as a routine part of studies evaluating diagnostics for WHO regulatory approval. Adding a survey at the end of an RCT or another quantitative study is not sufficient. Instead, the qualitative part could be embedded in a mixed method design or be a stand-alone study.

The general aim of these qualitative studies should be to examine:

- The experiences and challenges with diagnostic testing for TB (and HIV or other co-morbidities)
- Feasibility and acceptability of the new diagnostic in question or a similar diagnostic
- Values and preferences with regard to diagnosing TB and how new diagnostics change

these Such qualitative data will produce a whole range of potential issues the various users will have with a diagnostic technology and will point to possible uptake scenarios, potential pitfalls and barriers to utilization and access. These measures would create opportunities for meaningful engagement of users in WHO guideline development meetings, beyond the presence of one patient representative. It would allow gathering more diverse user perspectives and it could support defining additional PICO questions, for instance on operational dilemmas or ethical challenges for scale up.

13. Acknowledgements

We would like to thank all participants for the time and insights they provided.

14. References

Colvin, C. J. (2014). Anthropologies In and Of Evidence Making In Global Health Research and Policy. *Medical Anthropology, 34*(2), 99-105. doi:10.1080/01459740.2014.963196

Daftary, A. (2012). HIV and tuberculosis: the construction and management of double stigma. Soc Sci Med, 74(10), 1512-1519. doi:10.1016/j.socscimed.2012.01.027

Davids, M., Dheda, K., Pant Pai, N., Cogill, D., Pai, M., & Engel, N. (2015). A Survey on Use of Rapid Tests and Tuberculosis Diagnostic Practices by Primary Health Care Providers in South Africa: Implications for the Development of New Point-of-Care Tests. PLoS ONE, 10(10), e0141453. doi:10.1371/journal.pone.0141453

Engel, N., Davids, M., Blankvoort, N., Dheda, K., Pant Pai, N., & Pai, M. (2017). Making HIV testing work at the point of care in South Africa: a qualitative study of diagnostic practices. BMC Health Services Research, 17(1), 408. doi:10.1186/s12913-017-2353-6

Engel, N., Ganesh, G., Patil, M., Yellappa, V., Vadnais, C., Pai, N., & Pai, M. (2015). Point-of-care testing in India: missed opportunities to realize the true potential of point-of-care testing programs. BMC Health Services Research, 15(1), 550.

Engel, N., Yellappa, V., Davids, M., Dheda, K., Pai, N. P., & Pai, M. (2018). Barriers to Point of Care Testing in India and South Africa. In S. Hostettler, S. Najih Besson, & J.-C. Bolay (Eds.), Technologies for Development: From Innovation to Social Impact (pp. 75-85). Cham: Springer International Publishing.

McDowell, A., & Pai, M. (2016). Treatment as diagnosis and diagnosis as treatment: empirical

management of presumptive tuberculosis in India. INT J TUBERC LUNG DIS, 20(4), 536-543. doi:10.5588/ijtld.15.0562

McDowell, A., Raizada, N., Khaparde, S. D., Rao, R., Sarin, S., Kalra, A., . . . Denkinger, C. M. (2018). “Before Xpert I only had my expertise”: A qualitative study on the utilization and effects of Xpert technology among pediatricians in 4 Indian cities. PLoS ONE, 13(3), e0193656. doi:10.1371/journal.pone.0193656

Miller, R., Parkhurst, J. O., Peckham, S., & Singh, R. B. (2012). A qualitative study of HIV testing and referral practices of private hospital doctors treating patients with TB in Chennai, India. Int J Health Plann Manage, 27(2), 180-191.

Mvusi, L. (2018). Introduction of Urine LF-LAM Test in South Africa [PowerPoint slides]. Retrieved from

http://treatmentactiongroup.org/content/activists-guide-TB LAM-test

National AIDS and STI Control Programme. (2018). Guidelines on Use of Antiretroviral Drugs for Treating and Preventing HIV Infection in Kenya. Nairobi: Ministry of Health.

Shah, S. G. S., Robinson, I., & AlShawi, S. (2009). Developing medical device technologies from users' perspectives: A theoretical framework for involving users in the development process. International Journal of Technology Assessment in Health Care, 25(04), 514-521. doi:doi:10.1017/S0266462309990328

Squire, S. B., Belaye, A. K., Kashoti, A., Salaniponi, F. M., Mundy, C. J., Theobald, S., & Kemp, J. (2005). 'Lost' smear-positive pulmonary tuberculosis cases: where are they and why did we lose them? International Journal of Tuberculosis and Lung Disease, 9(1), 25-31.

TBCAB. (2018). Activists call on countries and donors to immediately scale up use of life-saving TB LAM test. TB Online. Retrieved from http://www.tbonline.info/posts/2018/10/25/activists-call- countries-and-donors-immediately/

Uganda National Tuberculosis and Leprosy Control Program. (2017). Manual for Management and

Control of tuberculosis and leprosy. Kampala: Ministry of Health.

World Health Oganization. (2012). Optimizing health worker roles to improve access to key maternal and newborn health interventions through task shifting. Retrieved from Geneva:

Yellappa, V., Devadasan, N., Krumeich, A., Pant Pai, N., Vadnais, C., Pai, M., & Engel, N. (2017). How patients navigate the diagnostic ecosystem in a fragmented health system: a qualitative study from India. *Global Health Action, 10*(1), 1350452. doi:10.1080/16549716.2017.1350452

Web Annex 4.15. Low complexity automated NAATs: Diagnostic accuracy for detection of resistance to isoniazid and second-line anti-TB agents. A systematic review

Samantha Pillay[1], Geraint R Davies[2], Marty Chaplin[3], Margaretha De Vos[4], Rob Warren[1], Karen R Steingart[5], Grant Theron[1]

[1]DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[2]Liverpool, UK
[3]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK
[4]FIND, Geneva, Switzerland
[5]Honorary Research Fellow, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

Corresponding author:
Grant Theron, PhD
gtheron@sun.ac.za

EXECUTIVE SUMMARY

Xpert® MTB/XDR Assay (Xpert MTB/XDR, Cepheid, Sunnyvale, USA) detects MTBC (Mycobacterium tuberculosis complex) DNA and genomic mutations associated with resistance to isoniazid, fluoroquinolones, ethionamide, and second-line injectable drugs (amikacin, kanamycin, capreomycin) in a single cartridge. Xpert MTB/XDR is intended for use as a reflex test in clinical specimens (unprocessed sputum or concentrated sputum sediments) already determined to be MTBC-positive. The test is included in a class of diagnostic technologies that are cartridge-based and of low complexity.

The proposed role for the test is to be used as an initial test for resistance to isoniazid and second-line drugs (replacement for line probe assays and pDST as initial tests). Favorable characteristics of Xpert MTB/XDR include rapidity (less than 90 minutes for a result), ease-of-use (same familiar process as Xpert MTB/RIF and Xpert Ultra), and detection of resistance directly in clinical specimens.

This systematic review summarizes the current literature on the accuracy of Xpert MTB/XDR for detection of resistance to isoniazid, fluoroquinolones, ethionamide, and amikacin as part of a World Health Organization process to develop guidelines for use of the test. This review does not include molecular drug susceptibility testing (DST) for kanamycin and capreomycin because, with the adoption of the new treatment regimens using all-oral medicines, the second-line injectable drugs are less relevant. We include molecular DST for amikacin because, of the second-line injectable drugs, amikacin is preferentially included on longer regimens when susceptibility has been demonstrated and adequate measures to monitor for adverse reactions can be ensured.

To identify studies, we searched multiple databases up to 6 September 2020 without language restriction. Two review authors independently assessed studies for eligibility. Two review authors independently extracted data from the included studies.

We stratified analyses by population, irrespective of rifampicin resistance and with detected rifampicin resistance, and target condition. We combined data using meta-analysis by fitting the bivariate random effects model. We performed all analyses stratified by type of reference standard, phenotypic DST (pDST), genotypic DST (gDST), and composite reference standard. For multicentre studies, we performed meta-analyses at the centre level (i.e. treating each centre as a separate study). We excluded MTBC-negative, MTBC-non-determinate, and inconclusive drug resistant results from analyses of diagnostic test accuracy. We performed sensitivity analyses by repeating the meta-analyses and excluding data from the manufacturer.

We identified three unpublished studies: Cepheid 2020, DIAMA 2020, and FIND 2020. All studies involved adults. One study evaluated archived frozen specimens and two studies evaluated sputum using a cross-sectional, prospective study design. The studies were in Benin, Cameroon, China, New Delhi, Moldova, Mumbai, and South Africa.

We did not identify any studies that assessed the accuracy of Xpert MTB/XDR for drug resistance in children.

As assessed by QUADAS-2, in the patient selection domain two studies were at low risk of bias and one study at unclear risk of bias because the manner of participant selection was not reported. In the reference standard domain, studies had low risk of bias for resistance to isoniazid, fluoroquinolones, and amikacin, and high risk of bias for resistance to ethionamide (for both pDST and gDST).

Xpert MTB/XDR for isoniazid resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity (95% confidence interval) were 94.2% (89.3 to 97.0) and 98.0% (95.2 to 99.2) (3 studies, 1605 participants, 61.9% with isoniazid resistance; high-certainty evidence for sensitivity and specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have isoniazid resistance, 66 would be Xpert MTB/XDR-positive: of these, 19 (29%) would not have isoniazid resistance (false-positives) and 934 would be Xpert MTB/XDR-negative: of these, 3 (0%) would have isoniazid resistance (false-negatives).

Xpert MTB/XDR for fluoroquinolone resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity were 93.1% (88.0 to 96.1) and 98.3% (94.5 to 99.5.) (3 studies, 1337 participants, 28.7.% with fluoroquinolone resistance; high-certainty evidence for sensitivity and moderate-certainty evidence for specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have fluoroquinolone resistance, 63 would be Xpert MTB/XDR-positive: of these, 16 (25%) would not have fluoroquinolone resistance (false-positives) and 937 would be Xpert MTB/XDR-negative: of these, 3 (0%) would have fluoroquinolone resistance (false-negatives).

Xpert MTB/XDR for ethionamide resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity were 56.6% (41.8 to 70.3) and 97.1% (91.9. to 99.0) (2 studies, 838 participants, 52.5% with ethionamide resistance; low-certainty evidence for sensitivity and moderate-certainty evidence for specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have ethionamide resistance, 56 would be Xpert MTB/XDR-positive: of these, 28 (50%) would not have ethionamide resistance (false-positives) and 944 would be Xpert MTB/XDR-negative: of these, 22 (2%) would have ethionamide resistance (false-negatives).

Xpert MTB/XDR for ethionamide resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, gDST
Xpert MTB/XDR pooled sensitivity and specificity were 96.4% (92.2 to 98.3) and 100.0% (82.5. to 100.0) (2 studies, 1001 participants, 28.0% with ethionamide resistance; moderate-certainty evidence for sensitivity and very low-certainty evidence for specificity).

Xpert MTB/XDR for amikacin resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity were 89.1% (80.9. to 94.1) and 99.5% (96.9 to 99.9) (2 studies, 1008 participants, 15.0% with amikacin resistance; high-certainty evidence for sensitivity and specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have amikacin resistance, 50 would be Xpert MTB/XDR-positive: of these, 5 (10%) would not have amikacin resistance (false-positives) and 950 would be Xpert MTB/XDR-negative: of these, 5 (1%) would have amikacin resistance (false-negatives).

For each drug, Xpert MTB/XDR pooled sensitivity and specificity estimates were similar in people irrespective of rifampicin resistance and people with detected rifampicin resistance. However, owing to enrolment criteria in the studies, we note that most participants were rifampicin resistant in all analyses.

The sensitivity analyses made little difference to any of the findings.

Authors' conclusions

- For resistance to isoniazid, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 94.2% against a reference standard of pDST.
- For resistance to fluoroquinolones, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 93.1% against a reference standard of pDST.
- For resistance to ethionamide, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 56.6% against a reference standard of pDST.
- For resistance to amikacin, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 89.1% against a reference standard of pDST.
- MTB/XDR specificity was > 97.0% in nearly all analyses.

The impact of Xpert MTB/XDR is expected to be affected by several factors, including the health care infrastructure, access to other diagnostic tests, the ability of the index test to detect tuberculosis (which is required for DST), and the prevalence of resistance to a given drug. Given that the test targets a limited number of resistance variants in specific genes, the test may perform differently in different settings. These results should, therefore, be interpreted with caution.

Future studies should assess the accuracy of Xpert MTB/XDR in different population groups, including children and people living with HIV. In addition, studies should assess the accuracy of Xpert MTB/XDR in different geological settings, in smear-negative specimens, and with different types of clinical specimens.

BACKGROUND

Early recognition and improved characterisation of tuberculosis drug resistance is a prerequisite for the rapid delivery of novel regimens to those who could benefit from them. For MDR/rifampicin-resistant-tuberculosis, the arrival of novel or repurposed drugs such as bedaquiline, clofazimine, and linezolid has revolutionized the efficacy of longer regimens, dispensing with the need for injectable drugs, and promising to deliver shorter all-oral regimens. Fluoroquinolones have an essential role and are also important for protecting second-line drugs like bedaquiline (WHO Consolidated Guidelines (Module 4) 2020).

While the availability of drug susceptibility testing using culture-based and molecular methods is increasing, coverage and availability of these technologies varies widely. For example, globally in 2019, only 59% of bacteriologically confirmed new tuberculosis cases were tested for rifampicin resistance. Among patients with rifampicin resistance, 71% were tested for resistance to fluoroquinolones, though coverage varied from around 35% in the Western Pacific to nearly 90% in Europe (WHO Global tuberculosis report 2020).

The development and scale-up of Xpert MTB/RIF was a major step toward improving tuberculosis and rifampicin resistance detection globally. The assay simultaneously tests for both conditions and offers a mostly automated hands-off solution deployable in many high tuberculosis burden settings. Xpert MTB/RIF has, however, been met with limitations. Of 48 high-burden countries,[5] only 18 countries (38%) reported that a WHO-recommended rapid diagnostic (which includes Xpert MTB/RIF) had been used as the initial test for more than half of their patients with tuberculosis (WHO Global tuberculosis report 2020).

The status quo for isoniazid susceptibility testing is worse. Although in high MDR-TB settings, the presence of rifampicin resistance alone has served as a proxy for MDR-TB and the basis for treatment decisions, emerging data suggest that, in some settings, rifampicin resistance testing has suboptimal specificity for MDR-TB (WHO Global tuberculosis report 2020). This means that testing for resistance to isoniazid (a critical first-line drug) is increasingly important. For instance, a study in the eastern Democratic Republic of the Congo found one in five rifampicin-resistant patients to be isoniazid susceptible when tested using the MTBDR*plus* line probe assay (Bismwa 2020), and the most recent South African National Survey of Drug Resistance found hotspots of rifampicin mono-resistance, where the prevalence ratio of such cases exceeded that of MDR-TB by as much as 30% (NICD 2016). Conversely, isoniazid resistance in the presence of rifampicin susceptibility (isoniazid mono-resistance) is also increasingly recognised as another emerging challenge in managing tuberculosis as it is an important enabler of MDR-TB (Sulis 2020).

Globally in 2019, 13% of new tuberculosis cases and 17% of previously treated tuberculosis cases had isoniazid resistance (WHO Global tuberculosis report 2020), yet isoniazid susceptibility testing is only generally done in patients who are rifampicin resistant. One reason for this is that genetic testing for isoniazid resistance is more complicated than testing for rifampicin resistance owing to a greater variety of resistance-associated variants (including large deletions) across several genes (including loci in *katG*, i*nhA*, and *ahpC*). Information on these mutations may not be routinely available in lower resource settings despite evidence showing that isoniazid resistance is associated with a three-fold increased risk of poor treatment outcomes (Espinal 2000) and hence should be treated with an intensified regimen including a fluoroquinolone (WHO Consolidated Guidelines (Module 4) 2020). Wider implementation of this modified regimen may reduce the risk of treatment failure and emergence of MDR-TB.

Though individualisation of MDR-TB treatment regimens according to susceptibility testing is promoted by guidance, gaps in infrastructure and personnel to support culture-based approaches may in part explain why, of an estimated 465,000 new cases of MDR/rifampicin-resistant-tuberculosis

[5] Forty-eight countries are in one or more of the three lists of high TB, TB/HIV and MDR-TB burden countries.

annually, only 44% were detected and notified (WHO Global tuberculosis report 2020). The WHO recommends that rapid techniques be used as the initial diagnostic tests to detect tuberculosis and rifampicin resistance in order to minimize delays in starting appropriate treatment (WHO Consolidated Guidelines (Module 3) 2020). The multiplexed nature of these new technologies theoretically permits susceptibility to be detected accurately and comprehensively for a single drug (where variants in multiple genes may cause resistance) and to several different drugs, each with their own sets of distinct resistance determinants. The flexibility of this technology offers the possibility of simultaneous detection of high confidence resistance causing mutations important for multiple drugs other than rifampicin.

This systematic review evaluated newly-developed rapid technologies that detect resistance to isoniazid, fluoroquinolones, ethionamide, and amikacin.

A glossary of terms is provided in Appendix 1.

Index test(s)

The index tests are rapid, cartridge-based nucleic acid amplification tests, of low complexity, for detection of resistance to isoniazid and second-line anti-tuberculosis drugs.

We define a *cartridge-based test* as one that may use single or multiple specimens and most reagents are enclosed in a disposable sealed container to which a clinical specimen is added. Almost all processes (such as DNA extraction and/or polymerase chain reaction (PCR) procedures) are performed within the container linked to the diagnostic platform. Cartridge-based tests may require an initial manual specimen treatment step prior to transfer of the material requiring testing into the cartridge.

Low complexity refers to a situation where no special infrastructure is required and basic laboratory skills are suitable to run the test, however, equipment may still be required.

Xpert® MTB/XDR Assay (Xpert MTB/XDR, Cepheid, Sunnyvale, USA) is the main index test in this review. Evidence on MeltPro® XDR-TB (MeltPro, Xiamen Zeesan Biotech Co., Ltd., China) provided by the manufacturer is summarized separately in Supplement A. No independent evaluations of MeltPro were identified.

Xpert MTB/XDR detects MTBC (*Mycobacterium tuberculosis* complex) DNA and genomic mutations associated with resistance to isoniazid, fluoroquinolones, ethionamide, and second-line injectable drugs (amikacin, kanamycin, capreomycin) in a single cartridge. This review does not include molecular drug susceptibility testing (DST) for kanamycin and capreomycin because, with the adoption of the new treatment regimens using all-oral medicines, the second-line injectable drugs are less relevant. We include molecular DST for amikacin because, of the second-line injectable drugs, amikacin is preferentially included in longer regimens when susceptibility has been demonstrated and adequate measures to monitor for adverse reactions can be ensured (Bainomugisa 2020; WHO Consolidated Guidelines (Module 4) 2020).

Xpert MTB/XDR is intended for use as a reflex test in specimens (unprocessed sputum or concentrated sputum sediments) determined to be MTBC-positive (Cepheid package insert 2020). The test could also be done on cultured isolates; however, this is not stated by the manufacturer as an intended use case. Several advantages of the assay are proposed.

- Faster time to result for molecular DST.
- Results in < 90 minutes.
- Same easy-to-use process as Xpert MTB/RIF Ultra.
- Run on existing GeneXpert® platforms equipped with 10-colour modules.

The limit of detection for *Mycobacterium tuberculosis* by Xpert MTB/XDR (136 CFU/mL in unprocessed sputum) (Cepheid package insert 2020) is similar to that of Xpert MTB/RIF (112.6 CFU/mL), but higher than that of Xpert Ultra (15.6 CFU/mL) (Chakravorty 2017). The manufacturer states that "Specimens with "MTB Trace DETECTED" results when tested with the Xpert MTB/RIF Ultra Assay are expected to be below the limit of detection of the MTB/XDR Assay and are not recommended for testing with the Xpert MTB/XDR Assay," (Cepheid package insert 2020). As with Xpert MTB/RIF and Xpert Ultra, Xpert MTB/XDR detects both live and dead bacteria (Cepheid report 2020).

The following information is from the Cepheid package insert (Cepheid package insert 2020).
- Regarding isoniazid, Xpert MTB/XDR bases detection of resistance on mutations in defined regions of the *katG* and *fabG1* genes, *oxyR-ahpC* intergenic region and *inhA* promoter region of the MTB genome.

- Regarding fluoroquinolones, Xpert MTB/XDR bases detection of resistance on mutations in the *gyrA* and g*yrB* quinolone resistance determining regions of the MTB genome.

- Regarding ethionamide, Xpert MTB/XDR bases detection of resistance on mutations in the i*nhA* promoter region of the MTB genome. In addition, it is noted that "mutations conferring ethionamide resistance are reported to be present in genomic regions not targeted by the Xpert MTB/XDR assay". Of interest, Brossier and colleagues found that 22/47 (47%) of ethionamide-resistant clinical isolates had mutations in *ethA*. Hence, the absence of mutations in the *inhA* promoter region does not preclude ethionamide resistance (Brossier 2011). Cepheid acknowledges that reporting ethionamide resistance based only on the detection of the *inhA* promoter mutations is a known limitation that may limit sensitivity though specificity may be unaffected.

- Regarding amikacin, Xpert MTB/XDR bases detection of resistance on mutations in a defined region of *rrs* of the MTB genome.

Table 1. Drug related gene targets, codon regions, and nucleotide sequences that determine presence of variants associated with drug resistance in the Xpert MTB/XDR assay

Drug	Gene target	Codon regions	Nucleotide
Isoniazid	*inhA* promoter (also used for tuberculosis detection)	not applicable	-1 to -32 intergenic region
	katG	311-319	939-957
	fabG1	199-210	597-630
	oxyR-ahpC intergenic region	not applicable	-5 to -50 intergenic region (or -47 to -92) *
Ethionamide	*inhA* promoter	not applicable	-1 to -32 intergenic region
Fluoroquinolones	*gyrA*	87-95	261-285
	gyrB	531-544 (or 493-505) *	1596-1632
Amikacin, Kanamycin, Capreomycin	*rrs*		1396-1417
	eis promoter	not applicable	-6 to -42 intergenic region

*Codon numbering system according to Camus JC, Pryor MJ, Médigue C, Cole ST. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiology (Reading). 2002;148(Pt 10):2967-2973, as reported in Cepheid, Clinical evaluation of the Xpert® MTBXDR assay, Report R244C2 Xpert MTB/XDR Rev 1.0.

Xpert MTB/XDR can report results as MTB NOT DETECTED or MTB DETECTED. If results are reported as MTB DETECTED, each drug is reported as resistance DETECTED or NOT DETECTED. If results are reported as MTB NOT DETECTED, INVALID, ERROR, or NO RESULT, then no DST results are reported, Appendix 2.

Clinical pathway

A clinical pathway presents a framework for developing recommendations about the use of a test and may assist in assessing the effect of a new test on management decisions and patient-important outcomes (Gopalakrishna 2016). We considered several clinical scenarios in Appendix 3,

In this systematic review, the intended use of Xpert MTB/XDR is for diagnosis of drug resistance. The role of the test would be a replacement test for culture-based phenotypic DST in people diagnosed with tuberculosis irrespective of rifampicin resistance or with detected rifampicin resistance.
The downstream consequences of testing include the following:

True-positive: people would benefit from rapid diagnosis and early initiation of appropriate tuberculosis treatment.

True-negative: people would be spared unnecessary treatment and would benefit from reassurance and pursuit of an alternative diagnosis.

False-positive: people would likely experience anxiety, morbidity from additional testing, possible delays in further diagnostic evaluation, and prolonged and unnecessary treatment with drugs that may have lower bactericidal activity than second-line regimens and often have serious adverse effects.

False-negatives: are at an increased risk of patient morbidity and mortality, and continued risk of community transmission of drug-resistant tuberculosis.

Review objective

To estimate the diagnostic accuracy of Xpert MTB/XDR on sputum for the diagnosis of the following conditions in people with microbiologically confirmed pulmonary tuberculosis.[6]

- Isoniazid resistance.
- Fluoroquinolone resistance.
- Ethionamide resistance.
- Amikacin resistance.

[6] We initially included an objective "to estimate the diagnostic accuracy of cartridge-based assays to diagnose pulmonary tuberculosis in people with signs and symptoms of pulmonary tuberculosis". However, we did not identify any studies that directly addressed this question. Therefore, this objective and the corresponding PICO questions were removed (Guideline Development Group Meeting, 1 November 2020), see Supplement B. The studies included in this review were designed to evaluate the manufacturer's intended use of Xpert MTB/XDR as a reflex test for a specimen (unprocessed sputum or concentrated sputum sediments) that is determined to be MTB positive (Cepheid package insert 2020).

METHODS

Types of studies

We included cross-sectional studies and cohort studies that assessed the diagnostic accuracy of the index test. We included diagnostic accuracy studies in which cases and controls were sampled from a single source population (referred to as a single gate design). We excluded case-control studies where cases and controls were sampled from different populations (referred to as a two-gate design). The latter type of study is prone to bias, particularly when a study enrols participants with severe disease and healthy participants without disease (Rutjes 2005). We included studies where the reference standard was performed after the index test and those where the reference standard was performed before the index test. We only included studies that reported data comparing the index test to an acceptable reference standard (defined below) from which we could extract true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

Participants

We included people of any age, HIV positive or negative, with microbiologically confirmed pulmonary tuberculosis. Participants with tuberculosis were included irrespective of rifampicin resistance (with or without rifampicin resistance, or rifampicin resistance unknown) or with detected rifampicin resistance. We included studies that assessed the diagnostic accuracy of the index test using sputum, consistent with the intended use of the manufacturer, and studies from all types of health facilities and all laboratory levels (peripheral, intermediate, and central) from all countries.

Index test

Xpert MTB/XDR is the main index test in this review. Evidence on MeltPro® XDR-TB (MeltPro, Xiamen Zeesan Biotech Co., Ltd., China) provided by the manufacturer is summarized separately in Supplement A.

Target conditions

We included four target conditions:

1. Isoniazid resistance.
2. Fluoroquinolone resistance.
3. Ethionamide resistance.
4. Amikacin resistance.

Reference standards

We included a microbiological reference standard (MRS) and a composite reference standard (CRS).

The microbiological reference standards were phenotypic DST (pDST) alone and genotypic DST (gDST) alone.

The composite reference standard was pDST and gDST, where at least one component test is positive.

In the methodological assessment using QUADAS-2, we took into account the reliability of these different reference standards for individual drugs (Heyckendorf 2017; WHO Critical concentrations 2018).

Outcomes

Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).
Cure - *we did not identify any studies that reported data for this outcome.*
Mortality - *we did not identify any studies that reported data for this outcome.*
Time to diagnosis - *we did not identify any studies that reported data for this outcome.*
Time to start treatment - *we did not identify any studies that reported data for this outcome.*

Search methods

We searched the following databases: Ovid MEDLINE (OVID, 1946-present) and Embase (OVID, 1947-present), for studies evaluating cartridge-based tests using tuberculosis, pulmonary AND Xpert, GeneXpert, Truenat, Cartridge, Point-of-Care Systems, Drug Susceptibility Test, isoniazid resistance, fluoroquinolone resistance, and second-line injectable drug resistance as search terms. We also searched Clinicaltrials.gov and the WHO ICTRP for trials in progress. Searches were run up to 6 September 2020 without language restriction, Appendix 4. On 4 November 2020, we ran an additional search using the search terms Zeesan and MeltPro.

We contacted researchers at FIND, the WHO Global Tuberculosis Programme, the manufacturer, and other experts in the field of tuberculosis diagnostics for information on ongoing and unpublished studies. We reviewed data submitted via the WHO public call.

Data collection and analysis

Selection of studies

We used Covidence to manage the selection of studies (Covidence 2017). Two review authors independently assessed studies for eligibility. We resolved disagreements by discussion with a third review author. We illustrated the study selection process in a PRISMA diagram (Moher 2009).

Data extraction

Two review authors independently extracted data from the reports, including: author, publication year, study design, country(ies)/sites where study was located, clinical setting, population characteristics, the number of TP, FP, FN, and TN values with respect to the reference standard, and inconclusive test results. We resolved disagreements by discussion with a third review author.

Assessment of methodological quality

Two review authors working independently assessed methodological quality using QUADAS-2 tailored to this review, Appendix 5. We resolved disagreements by discussion with a third review author.

Statistical analysis and data synthesis

We stratified analyses by population and target condition. Within each stratum, for example, detection of isoniazid resistance, we plotted estimates of the studies' observed sensitivities and specificities in forest plots with 95% confidence intervals (CIs) and in receiver operating characteristic (ROC) space using Review Manager (RevMan). Where adequate data were available, we combined data using meta-analysis by fitting the bivariate random effects model (Macaskill 2010; Reitsma 2005), using Stata with the metandi and xtmelogit commands (Stata 2019). When a bivariate random effects model could not be fit owing to few studies or sparse data, we instead specified two univariate random effects models (Takwoingi 2015). In situations where all studies in a meta-analysis reported a sensitivity of 100% or specificity of 100%, we used simple pooling by summing up the numbers of true positives and total resistant cases to calculate sensitivity or the numbers of true negatives and total susceptible cases to calculate specificity, as required. We performed all analyses stratified by type of reference standard.

For multicentre studies, we anticipated that there would be variability in terms of how laboratory practices were carried out between different centres. For this reason, when data were available, we performed meta-analyses at the centre level (i.e. treating each centre as a separate study).

We excluded MTBC-negative and inconclusive test results from analyses of diagnostic test accuracy.

Inconclusive index test results

The manufacturer defines two types of inconclusive results, non-determinate and indeterminate.

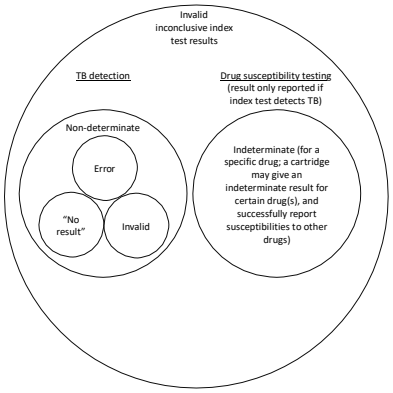


Figure. 1. Overview of different types of inconclusive results for Xpert MTB/XDR.

A non-determinate Xpert MTB/XDR test result is one that results in an INVALID, ERROR, or NO RESULT and can be due to an operator error, instrument, or cartridge issue (Cepheid package insert 2020). These three options are automatically generated, including the one called NO RESULT. The underlying reason for a non-determinate result is often not specified. The non-determinate Xpert MTB/XDR test results pertain only to the detection of MTBC, not to the detection of drug resistance.

A non-determinate result is distinct from MTB NOT DETECTED as shown below in Figure 2.

Drug Class	Result Call
N/A	INVALID/ERROR/NO RESULT
	MTB DETECTED
	MTB NOT DETECTED

Figure 2. Interpretation of non-determinate results and their relation to MTB DETECTED and MTB NOT Detected

An indeterminate Xpert MTB/XDR test result is one that indicates that resistance to a given drug could not definitively be detected based on the test's algorithm (Cepheid package insert 2020). This means that, based on quality control criteria, the test was not able to confidently report this particular result and the software suppressed the reporting of this. Indeterminate Xpert MTB/XDR test results pertain only to the detection of resistance to anti-tuberculosis drugs.

In addition, when data were available, we reported when the index test did not detect tuberculosis to begin with (missed cases).

We used the following approach to describe the different types of results.

Xpert MTB/XDR MTB NOT DETECTED
Among specimens with pDST results available, we determined the percentage that were Xpert MTB/XDR MTB NOT DETECTED. Among specimens with results reported as Xpert MTB/XDR

MTB NOT DETECTED, we further determined the percentage that were resistant or susceptible according to pDST.

Xpert MTB/XDR NON-DETERMINATE
Among the specimens initially tested, we determined the percentage of Xpert MTB/XDR NON-DETERMINATE results and, of these, the number of ERROR, INVALID, and NO RESULT results. We also determined the percentage of non-determinate results remaining following retesting.

Xpert MTB/XDR INDETERMINATE
Among specimens reporting Xpert MTB/XDR MTB DETECTED, we determined the percentage that were Xpert MTB/XDR INDETERMINATE (as drug resistance is only evaluated when MTB is detected). Among specimens with results reported as Xpert MTB/XDR INDETERMINATE, we further determined the percentage that were resistant or susceptible, according to pDST.

Investigations of heterogeneity
For each target condition, we investigated heterogeneity through visual examination of forest plots of sensitivity and specificity.

Sensitivity analyses
We performed sensitivity analyses by limiting inclusion in the meta-analysis to studies that were not designed or conducted by the manufacturer, therefore, we excluded Cepheid 2020.

Summary of findings and assessment of the certainty of the evidence (GRADE)
We assessed the certainty of evidence using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach for diagnostic studies (Schünemann 2008; Schünemann 2016). As recommended, we rated the certainty of evidence as either high (not downgraded), moderate (downgraded by one level), low (downgraded by two levels), or very low (downgraded by more than two levels) based on five domains: risk of bias, indirectness, inconsistency, imprecision, and publication bias. For each outcome, the certainty of evidence started as high for cross-sectional or cohort studies that enrolled participants with diagnostic uncertainty. When we found a reason for downgrading, we used our judgement to classify the reason as either serious (downgraded by one level) or very serious (downgraded by two levels). At least two review authors discussed judgments and applied GRADE in the following way (GRADEpro GDT 2015; Schünemann 2020a; Schünemann 2020b).

- Risk of bias: we used QUADAS-2 to assess risk of bias.
- Indirectness: we assessed indirectness in relation to the population (including disease spectrum), setting, interventions, and outcomes (accuracy measures). We used the prevalence of the condition as a guide to whether there was indirectness in the population.
- Inconsistency: GRADE recommends downgrading for unexplained inconsistency in sensitivity and specificity estimates.
- Imprecision: we considered a precise estimate to be one that would allow a clinically meaningful decision. We considered the width of the 95% CI and asked ourselves, ‘Would we make a different decision if the lower or upper boundary of the CI represented the truth?' In addition, we worked out projected ranges for TP, FN, TN, and FP for the prevalence of resistance to a given drug and made judgements on imprecision from these calculations.
- Publication bias: we considered the comprehensiveness of the literature search, outreach to researchers in tuberculosis, evidence identified from the WHO public call, and assistance from the WHO in identifying studies. Through these sources, we

identified several unpublished studies, but no publications. We graded publication bias as undetected.

RESULTS

Results of the search

We identified and screened a total of 1,649 records. Of these, we excluded 1620 for relevance to the topic. We retrieved 29 full text articles, including unpublished reports, and excluded 26 mainly because they were not rapid, low-complexity cartridge-based tests. We identified three unpublished studies for inclusion in the review, Cepheid 2020, DIAMA 2020, and FIND 2020. Appendix 6 shows the flow of studies in the review. A list of included and excluded studies is provided in Appendix 7.

Methodological quality of included studies

In the patient selection domain, we considered two studies (67%) to have low risk of bias and one study to have unclear risk of bias because we were unsure about the manner of participant selection (Cepheid 2020). Regarding applicability for patient selection, we considered all studies to have low concern.

In the index test domain, we considered all studies to have low risk of bias and low concern about applicability.

In the reference standard domain, we considered risk of bias separately for each drug and each reference standard. For resistance to isoniazid, fluoroquinolones, and amikacin, for pDST and gDST, we considered all studies have low risk of bias. For resistance to ethionamide, we considered all studies to have high risk of bias. For pDST, this was owing to considerable overlap in the minimum inhibitory concentration (MIC)s of *M tuberculosis* isolates with and without resistance-causing variants. For gDST, this was because no study included all loci required, e*thA*, *ethR*, and *inhA* promoter. Regarding applicability, for the reference standard domain, we considered all studies to have low concern.

In the flow and timing domain, we considered two studies to have low risk of bias and one study to have high risk of bias because not all participants were included in the analysis (DIAMA 2020). A summary table showing risk of bias and applicability concerns is included with each PICO question.

Findings

Study characteristics

The studies were in Benin, Cameroon, China, New Delhi, Moldova, Mumbai, Rwanda, and South Africa. We present key characteristics of the included studies in the Characteristics of included studies table, Appendix 8.

PICO questions

Should MTB/XDR assay on sputum be used to diagnose isoniazid resistance in patients with microbiologically confirmed pulmonary TB?

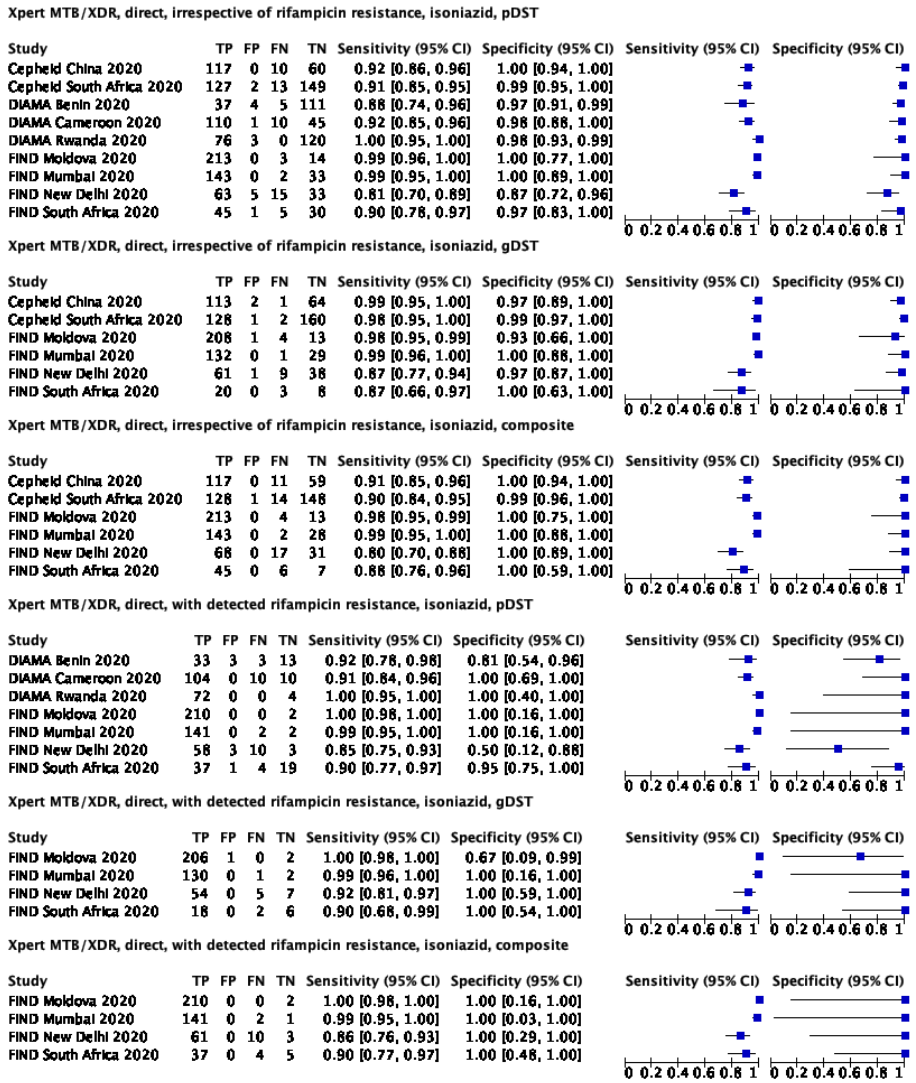


Figure 3. Forest plots of Xpert MTB/XDR sensitivity and specificity for detection of isoniazid resistance, by population and reference standard. Direct refers to testing directly on sputum.

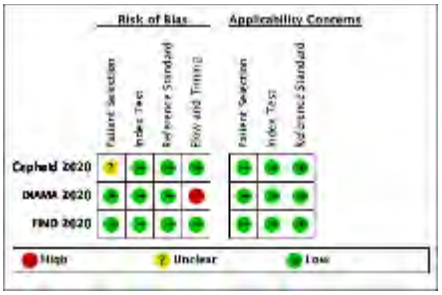


Figure 4. Xpert MTB/XDR, isoniazid resistance, risk of bias and applicability concerns.

Should MTB/XDR assay on sputum be used to diagnose fluoroquinolone resistance in patients with microbiologically confirmed pulmonary TB?

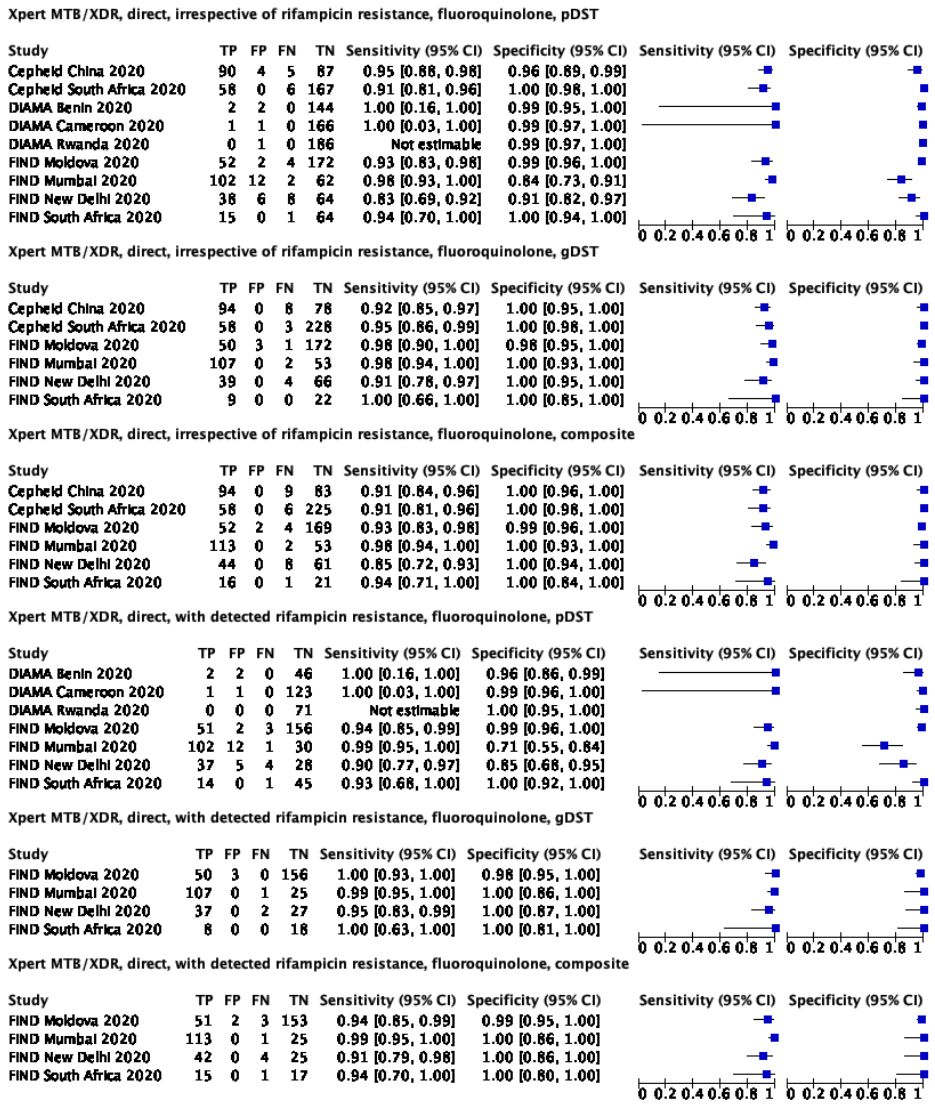


Figure 5. Forest plots of MTB/XDR sensitivity and specificity for detection of resistance to fluoroquinolones by population and reference standard. Direct refers to testing directly on sputum.

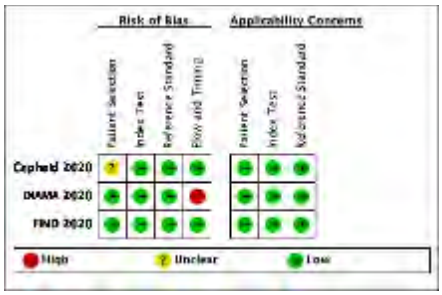


Figure 6. Xpert MTB/XDR, fluoroquinolone resistance, risk of bias and applicability concerns.

Should MTB/XDR assay on sputum be used to diagnose ethionamide resistance in patients with microbiologically confirmed pulmonary TB?

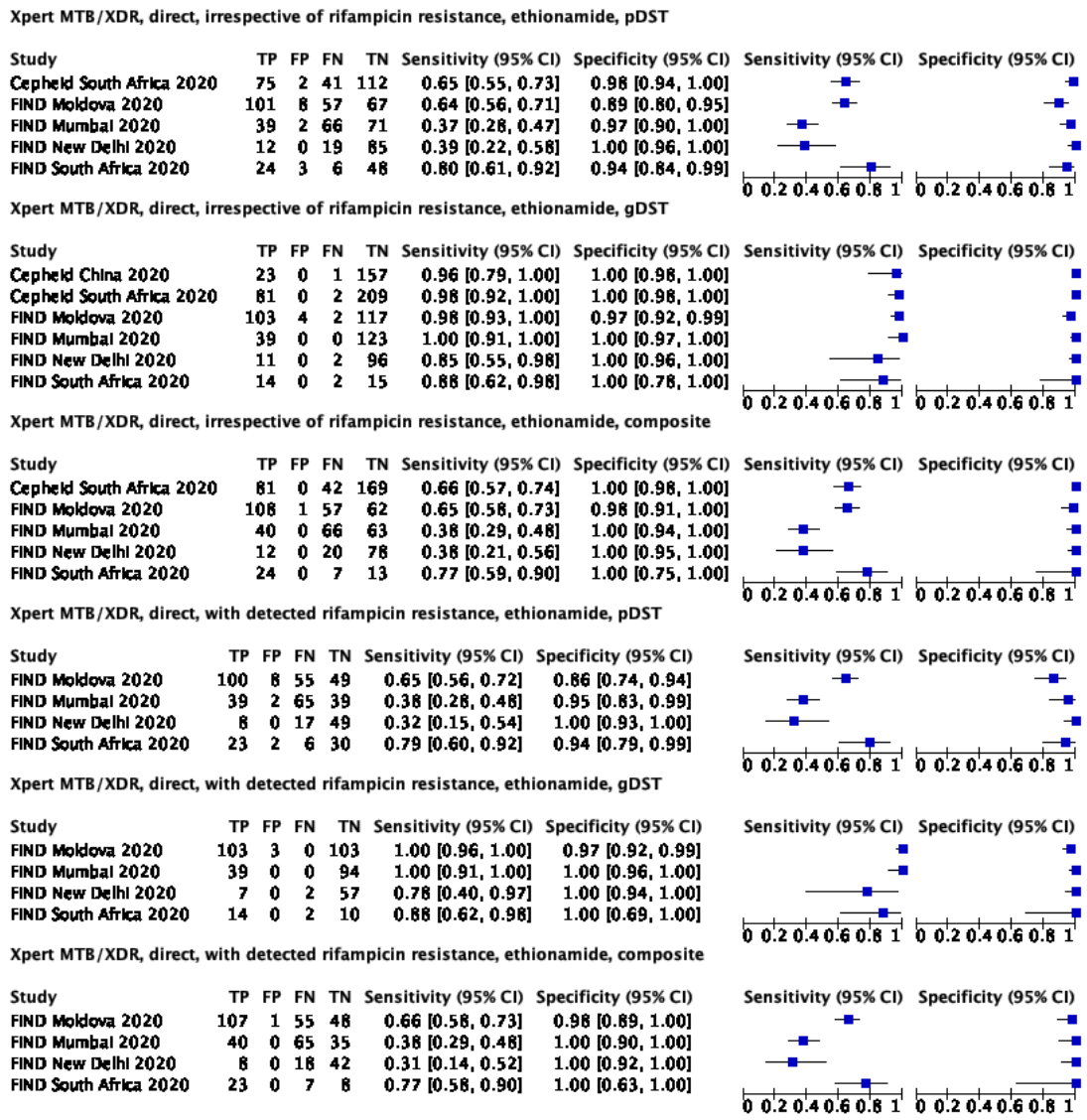


Figure 7. Forest plots of MTB/XDR sensitivity and specificity for detection of resistance to ethionamide by population and reference standard. Direct *refers to testing directly on sputum.*

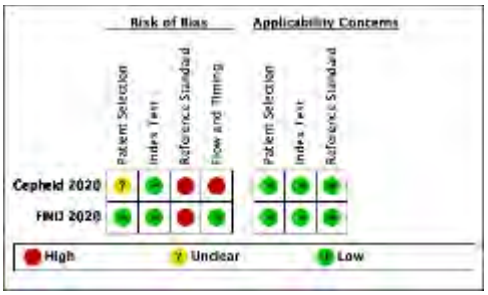


Figure 8. Xpert MTB/XDR, ethionamide resistance, risk of bias and applicability concerns.

Should MTB/XDR assay on sputum be used to diagnose amikacin resistance in patients with microbiologically confirmed pulmonary TB?

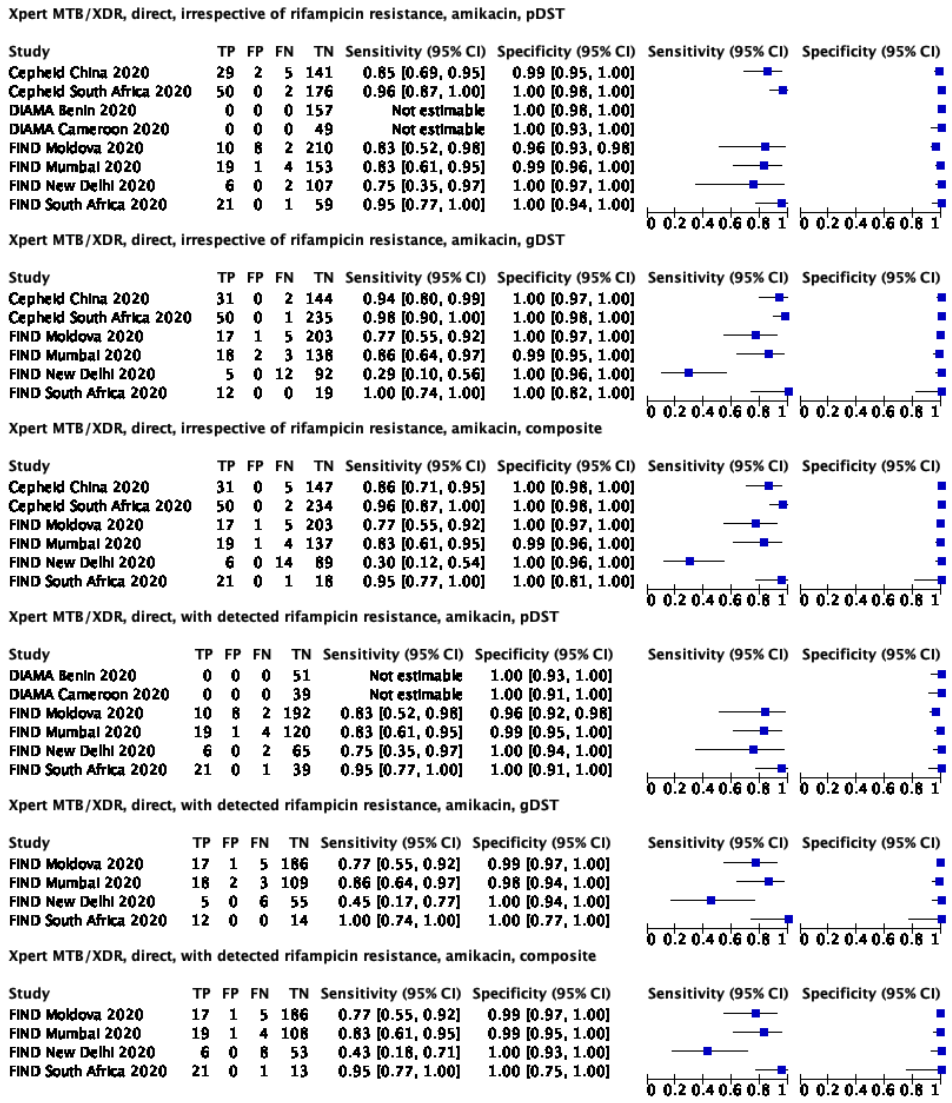


Figure 9. Forest plots of MTB/XDR sensitivity and specificity for detection of resistance to amikacin by population and reference standard. Direct refers to testing directly on sputum.

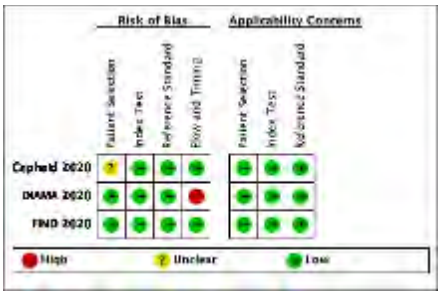


Figure 10. Xpert MTB/XDR, amikacin resistance, risk of bias and applicability concerns.

Table 2. Performance of Xpert MTB/XDR for resistance to isoniazid, fluoroquinolones, ethionamide, and amikacin

Drug	Reference standard	No. studies (participants)	No. (%) with drug resistance	Pooled sensitivity % (95% CI)	Pooled specificity % (95% CI)	Positive predictive value % (95% CI) [1]	Negative predictive value % (95% CI) [1]

Irrespective of rifampicin resistance							
INH	pDST	3 (1605)	994 (61.9)	94.2 (89.3 to 97.0)	98.0 (95.2 to 99.2)	71.3 (50.1 to 86.0)	99.7 (99.4 to 99.8)
INH	gDST	2 (999)	682 (68.3)	97.3 (92.8 to 99.0)	98.4 (95.9 to 99.3)	75.6 (55.4 to 88.6)	99.9 (99.6 to 100.0)
INH	Composite	2 (1055)	768 (72.8)	93.6 (86.5 to 97.1)	99.7 (96.6 to 100.0)	94.2 (58.6 to 99.5)	99.7 (99.3 to 99.8)
With detected rifampicin resistance							
INH	pDST	2 (744)	684 (91.9)	97.2 (89.7 to 99.3)	91.5 (68.5 to 98.1)	83.0 (51.2 to 95.8)	99.1 (96.6 to 99.8)
INH	gDST	1 (434)	416 (95.9)	98.4 (88.9 to 99.8)	97.5 (27.1 to 100.0)	94.5 (15.4 to 99.9)	99.5 (96.6 to 99.9)
INH	Composite	1 (476)	465 (97.7)	97.6 (84.7 to 99.7)	100.0 (74.1 to 100.0)	100.0 (58.0 to 100.0)	99.3 (95.2 to 99.9)
Irrespective of rifampicin resistance							
FQ	pDST	3 (1337)	384 (28.7)	93.1 (88.0 to 96.1)	98.3 (94.5 to 99.5)	74.6 (46.8 to 90.7)	99.7 (99.4 to 99.8)
FQ	gDST	2 (997)	375 (37.6)	95.7 (91.8 to 97.7)	99.9 (92.0 to 100.0)	97.5 (36.9 to 100.0)	99.8 (99.6 to 99.9)
FQ	Composite	2 (1021)	407 (39.9)	92.8 (88.1 to 95.8)	99.8 (96.0 to 100.0)	95.5 (54.4 to 99.7)	99.6 (99.4 to 99.8)
With detected rifampicin resistance							
FQ	pDST	2 (666)	216 (32.4)	95.2 (89.1 to 98.0)	96.6 (87.2 to 99.2)	92.4 (75.4 to 97.9)	98.5 (96.7 to 99.4)
FQ	gDST	1 (434)	205 (47.2)	98.6 (94.3 to 99.7)	98.8 (94.7 to 99.7)	97.2 (88.6 to 99.4)	99.6 (98.2 to 99.9)
FQ	Composite	1 (452)	230 (50.9)	96.0 (90.6 to 98.4)	99.1 (96.2 to 99.8)	97.9 (91.3 to 99.5)	98.8 (97.2 to 99.5)
Irrespective of rifampicin resistance							
ETO	pDST	2 (838)	440 (52.5)	56.6 (41.8 to 70.3)	97.1 (91.9 to 99.0)	50.9 (28.6 to 72.8)	97.8 (97.0 to 98.4)
ETO	gDST	2 (1001)	280 (28.0)	96.4 (92.2 to 98.3)	100.0 (82.5 to 100.0)	99.6 (19.5 to 100)	96.5 (92.7 to 98.4)
ETO	Composite	2 (843)	457 (54.2)	57.1 (42.8 to 70.2)	99.8 (95.3 to 100.0)	94.7 (39.9 to 99.8)	97.9 (97.1 to 98.5)
With detected rifampicin resistance							
ETO	pDST	1 (492)	313 (63.6)	51.7 (33.1 to 69.8)	94.8 (84.8 to 98.3)	81.0 (62.2 to 91.7)	86.7 (81.9 to 90.4)
ETO	gDST	1 (434)	167 (38.5)	98.0 (74.2 to 99.9)	99.7 (83.5 to 100.0)	99.3 (68.6 to 100.0)	99.4 (91.2 to 100.0)
ETO	Composite	1 (457)	323 (70.7)	53.1 (34.7 to 70.7)	99.5 (87.0 to 100.0)	98.0 (63.9 to 99.9)	87.6 (82.6 to 91.3)
Irrespective of rifampicin resistance							
AMK	pDST	2 (1008)	151 (15.0)	89.1 (80.9 to 94.1)	99.5 (96.9 to 99.9)	90.1 (59.0 to 98.3)	99.5 (99 to 99.7)
AMK	gDST	2 (990)	156 (15.8)	89.5 (64.5 to 97.6)	99.7 (98.4 to 99.9)	93.3 (73.9 to 98.6)	99.5 (97.9 to 99.9)
AMK	Composite	2 (1005)	175 (17.4)	84.1 (63.0 to 94.3)	99.8 (99.0 to 99.9)	94.9 (81.1 to 98.8)	99.2 (98 to 99.7)
With detected rifampicin resistance							
AMK	pDST	1 (490)	65 (13.3)	86.1 (75.0 to 92.7)	98.9 (93.0 to 99.8)	97.2 (83.4 to 99.6)	95.9 (92.7 to 97.8)
AMK	gDST	1 (433)	66 (15.2)	81.1 (56.0 to 93.6)	99.2 (96.9 to 99.8)	97.8 (92.4 to 99.4)	94.6 (86.8 to 97.9)
AMK	Composite	1 (443)	81 (18.3)	79.0 (55.4 to 91.9)	99.5 (97.6 to 99.9)	98.4 (93.7 to 99.6)	94.0 (86.8 to 97.4)

Abbreviations: AMK: amikacin; CI: Confidence interval; standard; DST: drug susceptibility testing; ETO: ethionamide; FQ: fluoroquinolone; INH: isoniazid; pDST phenotypic DST; gDST: genotypic DST.

Notes: Within each multicentre study, when data were available, we performed meta-analyses at the centre level (i.e. treating each centre separately).
1. Prevalence for calculating predictive values: 5% in people irrespective of rifampicin resistance and 30% in people with detected rifampicin resistance.

As seen in Table 2, for each drug, Xpert MTB/XDR pooled sensitivity and specificity estimates were similar in people irrespective of rifampicin resistance and people with detected rifampicin resistance. However, owing to enrolment criteria in the studies, we note that most participants were rifampicin resistant in all analyses.

PICO questions

1. Should MTB/XDR assay on sputum be used to diagnose INH resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, MRS?
2. Should MTB/XDR assay on sputum be used to diagnose INH resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, CRS?
3. Should MTB/XDR assay on sputum be used to diagnose INH resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, MRS?
4. Should MTB/XDR assay on sputum be used to diagnose INH resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, CRS?
5. Should MTB/XDR assay on sputum be used to diagnose FQ resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, MRS?
6. Should MTB/XDR assay on sputum be used to diagnose FQ resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, CRS?
7. Should MTB/XDR assay on sputum be used to diagnose FQ resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, MRS?
8. Should MTB/XDR assay on sputum be used to diagnose FQ resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, CRS?
9. Should MTB/XDR assay on sputum be used to diagnose ETO resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, pDST?
10. Should MTB/XDR assay on sputum be used to diagnose ETO resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, gDST?
11. Should MTB/XDR assay on sputum be used to diagnose ETO resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, CRS?
12. Should MTB/XDR assay on sputum be used to diagnose ETO resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, pDST?
13. Should MTB/XDR assay on sputum be used to diagnose ETO resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, gDST?
14. Should MTB/XDR assay on sputum be used to diagnose ETO resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, CRS?
15. Should MTB/XDR assay on sputum be used to diagnose AMK resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, MRS?
16. Should MTB/XDR assay on sputum be used to diagnose AMK resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, CRS?
17. Should MTB/XDR assay on sputum be used to diagnose AMK resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, MRS?
18. Should MTB/XDR assay on sputum be used to diagnose AMK resistance in patients with microbiologically confirmed pulmonary TB, with detected resistance to RIF, CRS?

Table 3. GRADE Certainty of Evidence
See Supplement C. GRADE evidence profiles.

PICO	Drug	Population	Reference standard	No. studies (participants)	Pooled sensitivity % (95% CI)	Pooled specificity % (95% CI)	Certainty Evidence Sens	Certainty Evidence Spec	Explanations
1	INH	Irrespective rifampicin resistance	pDST	3 (1605)	94.2 (89.3, 97.0)	98.0 (95.2, 99.2)	Moderate	Moderate	Downgraded one level for indirectness for sensitivity and specificity
2	INH	Irrespective rifampicin resistance	CRS	2 (1055)	93.6 (86.5, 97.1)	99.7 (96.6, 100.0)	Moderate	Moderate	Downgraded one level for indirectness for sensitivity and specificity
3	INH	With detected rifampicin resistance	pDST	2 (744)	97.2 (89.7, 99.3)	91.5 (68.5, 98.1)	High	Low	Downgraded one level for inconsistency, and one level imprecision (specificity)
4	INH	With detected rifampicin resistance	CRS	1 (476)	97.6 (84.7, 99.7)	100.0 (74.1 100.0)	High	Low	Downgraded two levels for imprecision (specificity)
5	FQ	Irrespective rifampicin resistance	pDST	3 (1337)	93.1 (88.0, 96.1)	98.3 (94.5, 99.5)	High	Moderate	Downgraded one level for inconsistency (specificity)
6	FQ	Irrespective rifampicin resistance	CRS	2 (1021)	96.0 (90.6, 98.4)	99.1 (96.2, 99.8)	High	High	
7	FQ	With detected rifampicin resistance	pDST	2 (666)	95.2 (89.1, 98.0)	96.6 (87.2, 99.2)	High	Moderate	Downgraded one level for inconsistency (specificity)
8	FQ	With detected rifampicin resistance	CRS	1 (452)	96.0 (90.6 to 98.4)	99.1 (96.2 to 99.8)	High	High	
9	ETO	Irrespective rifampicin resistance	pDST	2 (838)	56.6 (41.8, 70.3)	97.1 (91.9, 99.0)	Low	Moderate	Downgraded one level for risk of bias, one level for inconsistency (sensitivity); downgraded one level for risk of bias (specificity)
10	ETO	Irrespective rifampicin resistance	gDST	2 (1001)	96.4 (92.2, 98.3)	100.0 (82.5, 100.0)	Low	Very Low	Downgraded two levels for risk of bias (sensitivity); downgraded two levels for risk of bias, one level for imprecision (specificity)
11	ETO	Irrespective rifampicin resistance	CRS	2 (843)	57.1 (42.8, 70.2)	99.8 (95.3, 100.0)	Low	Moderate	Downgraded one level for risk of bias one level for inconsistency (sensitivity); downgraded one

									level for risk of bias (specificity)
12	ETO	With detected rifampicin resistance	pDST	1 (492)	51.7 (33.1, 69.8)	94.8 (84.8, 98.3)	Very Low	Moderate	Downgraded one level for risk of bias, one level for inconsistency, one level for imprecision (sensitivity); downgraded one level for risk of bias (specificity)
13	ETO	With detected rifampicin resistance	gDST	1 (434)	98.0 (74.2, 99.9)	99.7 (83.5, 100.0)	Very Low	Very Low	Downgraded two levels for risk of bias, one level for imprecision (sensitivity and (specificity)
14	ETO	With detected rifampicin resistance	CRS	1 (457)	53.1 (34.7, 70.7)	99.5 (87.0, 100.0)	Very Low	Moderate	Downgraded one level for risk of bias, one level for inconsistency, one level for imprecision (sensitivity); downgraded one level for risk of bias (specificity)
15	AMK	Irrespective rifampicin resistance	pDST	2 (1008)	89.1 (80.9, 94.1)	99.5 (96.9, 99.9)	Moderate	High	Downgraded one level for risk of bias (sensitivity)
16	AMK	Irrespective rifampicin resistance	CRS	2 (1005)	84.1 (63.0, 94.3)	99.8 (99.0, t99.9)	Low	High	Downgraded one level for risk of bias, one level for inconsistency (sensitivity)
17	AMK	With detected rifampicin resistance	pDST	1 (490)	86.1 (75.0, 92.7)	98.9 (93.0, 99.8)	Low	High	Downgraded two levels for imprecision (sensitivity)
18	AMK	With detected rifampicin resistance	CRS	1 (443)	79.0 (55.4, 91.9)	99.5 (97.6, 99.9)	Low	High	Downgraded two levels for imprecision (sensitivity)

Abbreviations: AMK: amikacin; CI: Confidence interval; standard; DST: drug susceptibility testing; ETO: ethionamide; FQ: fluoroquinolone; INH: isoniazid; pDST phenotypic DST; gDST: genotypic DST

Xpert MTB/XDR MTB NOT DETECTED and inconclusive test results

Xpert MTB/XDR MTB NOT DETECTED
Here we summarize results for Xpert MTB/XDR MTB NOT DETECTED and resistant cases therefore missed. Cepheid 2020 was the only study that reported this information.

Isoniazid

Of 530 specimens tested, 512 had pDST results available. Of these 512 specimens with pDST results available, 32 (6.3%) were Xpert MTB/XDR MTB NOT DETECTED.
By the pDST reference standard, of these 32 specimens, two (6.3%) were resistant and 30 (93.8%) were susceptible.

Fluoroquinolones

Of 530 specimens tested, 453 had pDST results available. Of these 453 specimens with pDST results available, 32 (7.1%), were Xpert MTB/XDR MTB NOT DETECTED.
By the pDST reference standard, of these 32 specimens, one (3.1%) was resistant and 31 (96.9%) were susceptible.

Ethionamide

Of 530 specimens tested, 260 had pDST results available. Of these 260 specimens with pDST results available, 30 (11.5%) were Xpert MTB/XDR MTB NOT DETECTED.
By the pDST reference standard, of these 30 specimens, two (6.7%) were resistant and 28 (93.3%) were susceptible.

Amikacin

Of 530 specimens tested, 445 had pDST results available. Of these 445 specimens, 32 (7.2%) were Xpert MTB/XDR MTB NOT DETECTED.
By the pDST reference standard, of these 32 specimens, 32 (100.0%) were susceptible.

Non-determinate test results
Here we provide a summary of non-determinate results and their pDST status.

Cepheid 2020
- Initial testing

Of 531 specimens tested, 15 resulted in non-determinate results after their Xpert testing. There were 10 “Error” results, one "Invalid” result, and four “No Result” results. *Therefore, the non-determinate rate upon initial testing was 2.8%.*
- Retesting

These 15 specimens were retested and 14 of the 15 gave valid results upon retest. One of the 15 retested specimens resulted in an “Error” following its repeat test. *Therefore, the non-determinate rate following retesting was 0.2% (1/531).*

FIND 2020
- Initial testing

Of 709 specimens tested, 21 resulted in non-determinate results after their initial Xpert tests. *Therefore, the non-determinate rate upon initial testing rate was 3.0% (21/709).*
- Retesting

Of these 21 specimens, 19 gave valid results upon retesting. *Therefore, the non-determinate rate following retesting was 0.3% (2/709).*

The phenotypic status of non-determinate results was not discernable for either study.

Indeterminate test results
Here we provide a summary of indeterminate results and their pDST status.

Isoniazid

Cepheid 2020
Of 530 specimens tested, 498 specimens had an Xpert MTB/XDR MTB DETECTED result. Of these 498 specimens, two (0.4%) had indeterminate results for detection of resistance.
By the pDST reference standard, of these two specimens, two (100%) were resistant and zero (0%) were susceptible.

FIND 2020
Of 709 specimens tested, 657 had an Xpert MTB/XDR MTB DETECTED result. Of these 657 specimens, two (0.3%) had indeterminate results for detection of resistance. None were indeterminate upon retesting.

Fluoroquinolones

Cepheid 2020
Of 530 specimens tested, 498 specimens had an Xpert MTB/XDR MTB DETECTED result. Of these 498 specimens, four (0.8%) had indeterminate results for detection of resistance.
By the pDST reference standard, of these four specimens, zero (0%) were resistant and four (100%) were susceptible.

FIND 2020
Of 709 specimens tested, 657 had an Xpert MTB/XDR MTB DETECTED result. Of these 657 specimens, nine (1.4%) had indeterminate results for detection of resistance. None were indeterminate upon retesting.

Ethionamide

Cepheid 2020
Of 530 specimens tested, 498 specimens had an Xpert MTB/XDR MTB DETECTED result. Of these 498 specimens, none (0%) had an indeterminate result for detection of resistance.

FIND 2020
Of 709 specimens tested, 657 had an Xpert MTB/XDR MTB Detected result. Of these 657 specimens, one (0.2%) had an indeterminate result for detection of resistance. This specimen was no longer indeterminate upon retesting.

Amikacin

Cepheid 2020
Of 530 specimens tested, 498 specimens had an Xpert MTB/XDR MTB DETECTED result. Of these 498 specimens, eight (1.6%) had indeterminate results for detection of resistance. By the pDST reference standard, of these eight specimens, zero (0%) were resistant and eight (100%) were susceptible.

FIND 2020
Of 709 specimens tested, 657 had an Xpert MTB/XDR MTB DETECTED result. Of these 657 specimens, 23 (3.5%) had indeterminate results for detection of resistance. One was indeterminate upon retesting.

pDST results could not be discerned for FIND 2020 indeterminates.

Sensitivity analyses

Table 4 presents the findings from sensitivity analyses that excluded data from the manufacturer. There are two rows of results presented for each drug. The first row presents the results of the meta-analysis including Cepheid 2020, and the subsequent row, the results of the sensitivity analysis excluding Cepheid 2020 (in bold).

These sensitivity analyses made little difference to any of the findings.

Table 4. Xpert MTB/XDR accuracy for drug resistance in people irrespective of rifampicin resistance, sensitivity analyses

Drug	Reference standard	No. studies (participants)	No. (%) with resistance to drug	Pooled sensitivity % (95% CI)	Pooled specificity % (95% CI)
Isoniazid	pDST	3 (1605)	994 (61.9)	94.2 (89.3 to 97.0)	98.0 (95.2 to 99.2)
Isoniazid, without Cepheid	**pDST**	**2 (1005)**	**685 (68.2)**	**96.0 (89.4 to 98.6)**	**97.1 (91.9 to 99.0)**
Fluoroquinolones	pDST	3 (1337)	384 (28.7)	93.1 (88.0 to 96.1)	98.3 (94.5 to 99.5)
Fluoroquinolones, without Cepheid	**pDST**	**2 (1112)**	**225 (20.1)**	**93.5 (83.4 to 97.6)**	**98.4 (94.3 to 99.5)**
Ethionamide	pDST	2 (838)	440 (52.5)	56.6 (41.8 to 70.3)	97.1 (91.9 to 99.0)
Ethionamide, without Cepheid	**pDST**	**1 (756)**	**324 (42.9)**	**53.1 (35.7 to 69.7)**	**96.5 (89.1 to 98.9)**
Amikacin	pDST	2 (1008)	151 (15.0)	89.1 (80.9 to 94.1)	99.5 (96.9 to 99.9)
Amikacin, without Cepheid	**pDST**	**1 (612)**	**65 (10.6)**	**86.1 (74.9 to 92.8)**	**99.3 (94.4 to 99.9)**

DISCUSSION

Summary of main results

This systematic review summarizes the current literature and included three unpublished studies on the accuracy of Xpert MTB/XDR for detection of resistance to isoniazid, fluoroquinolones, ethionamide, and amikacin.

Xpert MTB/XDR for isoniazid resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity (95% confidence interval) were 94.2% (89.3 to 97.0) and 98.0% (95.2 to 99.2) (3 studies, 1605 participants, 61.9% with isoniazid resistance; high-certainty evidence for sensitivity and specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have isoniazid resistance, 66 would be Xpert MTB/XDR-positive: of these, 19 (29%) would not have isoniazid resistance (false-positives) and 934 would be Xpert MTB/XDR-negative: of these, 3 (0%) would have isoniazid resistance (false-negatives).

Xpert MTB/XDR for fluoroquinolone resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity were 93.1% (88.0 to 96.1) and 98.3% (94.5 to 99.5.) (3 studies, 1337 participants, 28.7.% with fluoroquinolone resistance; high-certainty evidence for sensitivity and moderate-certainty evidence for specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have fluoroquinolone resistance, 63 would be Xpert MTB/XDR-positive: of these, 16 (25%) would not have fluoroquinolone resistance (false-positives) and 937 would be Xpert MTB/XDR-negative: of these, 3 (0%) would have fluoroquinolone resistance (false-negatives).

Xpert MTB/XDR for ethionamide resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST
Xpert MTB/XDR pooled sensitivity and specificity were 56.6% (41.8 to 70.3) and 97.1% (91.9. to 99.0) (2 studies, 838 participants, 52.5% with ethionamide resistance; low-certainty evidence for sensitivity and moderate-certainty evidence for specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have ethionamide resistance, 56 would be Xpert MTB/XDR-positive: of these, 28 (50%) would not have ethionamide resistance (false-positives) and 944 would be Xpert MTB/XDR-negative: of these, 22 (2%) would have ethionamide resistance (false-negatives).

Xpert MTB/XDR for ethionamide resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, gDST
Xpert MTB/XDR pooled sensitivity and specificity were 96.4% (92.2 to 98.3) and 100.0% (82.5. to 100.0) (2 studies, 1001 participants, 28.0% with ethionamide resistance; moderate-certainty evidence for sensitivity and very low-certainty evidence for specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have ethionamide resistance, 48 would be Xpert MTB/XDR-positive: of these, 0 (0%) would not have ethionamide resistance (false-positives) and 952 would be Xpert MTB/XDR-negative: of these, 2 (0%) would have ethionamide resistance (false-negatives).

Xpert MTB/XDR for amikacin resistance in people with microbiologically confirmed pulmonary tuberculosis irrespective of rifampicin resistance, pDST

Xpert MTB/XDR pooled sensitivity and specificity were 89.1% (80.9. to 94.1) and 99.5% (96.9 to 99.9) (2 studies, 1008 participants, 15.0% with amikacin resistance; high-certainty evidence for sensitivity and specificity).

Results of these studies indicate that in theory, of 1000 people where 50 have amikacin resistance, 50 would be Xpert MTB/XDR-positive: of these, 5 (10%) would not have amikacin resistance (false-positives) and 950 would be Xpert MTB/XDR-negative: of these, 5 (1%) would have amikacin resistance (false-negatives).

AUTHORS' CONCLUSIONS

- For resistance to isoniazid, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 94.2% against a reference standard of pDST.
- For resistance to fluoroquinolones, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 93.1% against a reference standard of pDST.
- For resistance to ethionamide, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 56.6% against a reference standard of pDST and 96.4% against a reference standard of gDST. However, the gDST reference standard only included the *inhA* promoter.
- For resistance to amikacin, in people irrespective of rifampicin resistance, Xpert MTB/XDR sensitivity was 89.1% against a reference standard of pDST.
- Xpert MTB/XDR specificity was > 97.0% in nearly all analyses.
- Overall, for resistance to isoniazid and fluoroquinoles, Xpert MTB/XDR sensitivity estimates for individual studies were consistent against the different reference standards.
- Overall, for resistance to a given drug, indeterminate results were infrequent and mostly resolved with retesting.
- We were not always able to link the analyses to a specific clinical pathway scenario, especially for Scenario A (patients evaluated for tuberculosis) and Scenario D (patients on treatment).

The impact of Xpert MTB/XDR is expected to be affected by several factors, including the health care infrastructure, access to other diagnostic tests, the ability of the index test to detect tuberculosis (which is required for DST), and the prevalence of resistance to a given drug. Given that the test targets a limited number of resistance variants in specific genes, the test may perform differently in different settings. These results should, therefore, be interpreted with caution.

The 2020 World Health Organization consolidated guidelines on drug resistant tuberculosis treatment recognize the importance of later generation fluoroquinolones in all-oral regimens of shorter duration (WHO Consolidated Guidelines (Module 4) 2020**)**. The review findings suggest that Xpert MTB/XDR provides accurate results for detection of fluoroquinolone resistance and can assist with rapid initiation of an optimized treatment regimen.

Future studies should assess the accuracy of Xpert MTB/XDR for drug resistance in different population groups, including children and people living with HIV. In addition, studies should assess the accuracy of Xpert MTB/XDR in different geological settings, in smear-negative specimens, and with different types of clinical specimens. Guidance is needed for specimens that test “MTB Trace DETECTED” with the Xpert MTB/RIF Ultra Assay.

Studies should utilize a comprehensive composite reference standard for gDST using all known resistance-associated loci, not just those analyzed by the index test. Studies should include patients from different points on the clinical pathway. In addition, we suggest quantifying the impact of non-actionable results, especially in smear-negative specimens. Future studies should also assess the

diagnostic accuracy of Xpert MTB/XDR for pulmonary tuberculosis in adults, children, and people living with HIV and in people who are smear negative.

Acknowledgements

We are grateful to Vittoria Lutje, Cochrane Infectious Diseases' Information Specialist, for help with the search strategy. We are grateful to people at Cepheid and FIND and to all members of the DIAMA consortium for providing data and answering our many questions.

Declarations of Interest

SP received funding from the World Health Organization Global Tuberculosis Programme, Geneva.
GRD has no conflicts to declare.
MDV is employed by the Foundation for Innovative New Diagnostics (FIND). FIND has conducted studies and published on Xpert MTB/RIF as part of a collaborative project between FIND, a Swiss non-profit, Cepheid, a US company, and academic partners. The product developed through this partnership was developed under a contract that obligated FIND to pay for development costs and trial costs and Cepheid to make the test available at specified preferential pricing to the public sector in low- and middle-income countries. In addition, FIND conducted studies for the Xpert MTB/RIF Ultra assay, which have also been published.
MC received funding from READ-It. READ-It aims to improve the evidence base and ensure its dissemination and helps to ensure healthcare problems relevant to low- and middle-income countries are addressed, and that people living in these countries are part of the process. READ-It (project number 300342-104) is funded by the Foreign, Commonwealth and Development Office (FCDO), UK.
RW has no conflicts to declare.
KRS received funding from the World Health Organization Global Tuberculosis Programme, Geneva. In addition, she has received financial support from Cochrane Infectious Diseases, Liverpool; McGill University, Montreal; Baylor College of Medicine, Houston; Stellenbosch University, Cape Town; Foundation for Innovative New Diagnostics (FIND), Geneva; and the World Health Organization Global Tuberculosis Programme, Geneva for preparation of related systematic reviews.
GT received funding from the World Health Organization Global Tuberculosis Programme, Geneva. In addition, he has received in-kind research consumable and equipment donations provided to employer by Cepheid to work on Xpert MTB/RIF and Xpert MTB/RIF Ultra (not Xpert MTB/XDR) for diagnostic accuracy evaluations for tuberculosis detection. These studies are on different products to those potentially considered for inclusion in this review.

References

Bainomugisa 2020
Bainomugisa A, Gilpin C, Coulter C, Marais BJ. New Xpert MTB/XDR: added value and future in the field. European Respiratory Journal 2020;56:2003616.
Bisimwa 2020
Bisimwa BC, Nachega JB, Warren RM, Theron G, Metcalfe JZ, Shah M, et al. Xpert MTB/RIF-detected rifampicin resistance is a sub-optimal surrogate for multidrug resistant tuberculosis in Eastern Democratic Republic of the Congo: diagnostic and clinical implications. Clinical Infectious Diseases 2020;ciaa873.
Brossier 2011
Brossier F, Veziris N, Truffot-Pernot C, Jarlier V, Sougakoff W. Molecular investigation of resistance to the antituberculous drug ethionamide in multidrug-resistant clinical isolates of Mycobacterium tuberculosis. Antimicrobial agents and chemotherapy 2011;55(1):355-60.
Cepheid package insert 2020
Cepheid. Xpert® MTB/XDR. GXMTB/XDR-10. Package insert 2020.
Chakravorty 2017
Chakravorty S, Simmons AM, Rowneki M, Parmar H, Cao Y, Ryan J, et al. The new Xpert MTB/RIF Ultra: improving detection of Mycobacterium tuberculosis and resistance to rifampin in an assay suitable for point-of-care testing. Molecular Biology 2017;8(4):e00812-17.
Covidence 2017
Covidence systematic review software. Available at www.covidence.org [Computer program]. Melbourne: Veritas Health Innovation, 2017.
GRADEpro GDT 2015

GRADEpro GDT [Computer program]. Version accessed 29 October 2016. Hamilton (ON): McMaster University (developed by Evidence Prime), 2015. Available at gradepro.org.

Gopalakrishna 2016

Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Defining the clinical pathway in Cochrane diagnostic test accuracy reviews. BMC Med Res Methodol. 2016;16(1):153.

Heyckendorf 2017

Heyckendorf J, Andres S, Köser CU, Olaru ID, Schön T, Sturegård E, et al. What is resistance? Impact of phenotypic versus molecular drug resistance testing on therapy for multi-and extensively drug-resistant tuberculosis. Antimicrobial Agents and Chemotherapy 2018;62(2):e01550-17.

Kohli 2020

Kohli M, MacLean E, Pai M, Schumacher SG, Denkinger, CM. Diagnostic accuracy of centralised assays for TB detection and detection of resistance to rifampicin and isoniazid: A systematic review and meta-analysis. European Respiratory Journal 2020.

Macaskill 2010

Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editor(s). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from: srdta.cochrane.org/.

Moher 2009

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. PLOS Medicine 6;7:e1000097. [DOI: 10.1371/journal.pmed1000097]

Reitsma 2005

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005;58(10):982-90.

Rutjes 2005

Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005 Aug;51(8):1335-41

Schünemann 2008

Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336(7653):1106-10.

Schünemann 2016

Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al; GRADE Working Group. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. Journal of Clinical Epidemiology 2016;76:89-98. [DOI: 10.1016/j.jclinepi.2016.01.032]

Schünemann 2020a

Schünemann HJ, Mustafa R, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy. Journal of Clinical Epidemiology 2020;122:129-41. [DOI: 10.1016/j.jclinepi.2019.12.020]

Schünemann 2020b

Schünemann HJ, Mustafa R, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. Inconsistency, Imprecision, publication bias and other domains for rating the certainty of evidence for test accuracy and presenting it in evidence profiles and summary of findings tables. Journal of Clinical Epidemiology 2020;122:142-52. [DOI: 10.1016/j.jclinepi.2019.12.021]

Sulis 2020

Sulis G, Pai M. Isoniazid-resistant tuberculosis: A problem we can no longer ignore. PLOS Medicine. 2020 17(1):e1003023

Takwoingi 2015

Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Statistical Methods in Medical Research 2015;26(4):1896-1911.Whiting 2011

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. Annals of Internal Medicine 2011;155(8):529-36.

WHO Consolidated Guidelines (Module 3) 2020

World Health Organization. WHO consolidated guidelines on tuberculosis. Module 3: diagnosis – rapid diagnostics for tuberculosis detection; June 2020. who.int/publications/i/item/who-consolidated-guidelines-on-tuberculosis-module-3-diagnosis---rapid-diagnostics-for-tuberculosis-detection (accessed 1 July 2020).

WHO Consolidated Guidelines (Module 4) 2020

World Health Organization. Global tuberculosis report 2020. who.int/tb/publications/global_report/en/ (accessed 19 October 2020).

This report has been prepared for the WHO Global TB Programme. Please do not distribute further.

WHO Critical concentrations 2018
World Health Organization. Technical Report on critical concentrations for drug susceptibility testing of medicines used in the treatment of drug-resistant tuberculosis. (WHO/CDS/TB/2018.5). License: CC BY-NC-SA 3.0 IGO. who.int/tb/publications/2018/WHO_technical_report_concentrations_TB_drug_susceptibility/en/ (accessed 16 July 2020).
WHO Definitions and reporting 2020
World Health Organization. Definitions and reporting framework for tuberculosis – 2013 revision (updated December 2014 and January 2020). who.int/tb/publications/definitions/en/ (accessed 24 July 2020).

APPENDICES

Appendix 1. Glossary of terms

Amplification
Amplification is replication of a DNA fragment to generate copies. Both the original and the newly synthesized copies can be described as the amplicons.

Codon
A codon is a sequence of three DNA or RNA bases that corresponds to a specific amino acid or a signal to start or stop transcription or translation. The DNA in coding regions of the genome is read in groups of three bases (A, G, C, T).

Critical concentration
The critical concentration of an anti-tuberculous agent has been adopted and modified from international convention. The critical concentration is defined as the lowest concentration of an anti-tuberculosis agent *in vitro* that will inhibit the growth of 99% of phenotypically wild type strains of *M tuberculosis* complex.

Culture isolate
Culture isolate refers to *M tuberculosis* cells from a clinical specimen that have been grown. For tuberculosis diagnosis, a volume of the clinical specimen is processed and incubated under conditions that promote *M tuberculosis* growth. The cells that are grown are referred to a culture isolate.

DNA sequencing
DNA sequencing is a process to determine the nucleotide (A, G, C, T) sequence of fragments of DNA. By comparison of DNA sequences from distinct tuberculosis isolates, variations known as mutations can be identified. Some mutations in *M tuberculosis* are known to be associated with drug resistance.

Drug susceptibility testing
Drug susceptibility tests determine whether *M tuberculosis* cells are sensitive or resistant to antibiotics. Testing may be undertaken using phenotypic or genotypic analyses.

eis promoter
Gene target included in the Xpert MTB/XDR test to detect mutations that confer resistance to second line injectable drugs, amikacin and kanamycin.

fabG1
Gene target included in the Xpert MTB/XDR test to detect mutations that confer resistance to isoniazid.

Genotypic drug susceptibility testing (gDST)
Genotypic testing involves detecting predetermined mutations in DNA that are known to make the organism resistant to a drug. When mutations causing drug resistance are not known, genotypic DST is not useful.

gyrA
Gene target included in the Xpert MTB/XDR test to detect mutations that confer resistance to fluoroquinolones.

gyrB
Gene target included in the Xpert MTB/XDR test to detect mutations that confer resistance to fluoroquinolones.

Heteroresistance

Heteroresistance is defined as resistance to certain antibiotics in a subset of a larger microbial population that is generally considered to be susceptible to these antibiotics according to traditional phenotypic drug susceptibility testing.

Indeterminate test result

An indeterminate Xpert MTB/XDR test result is one that indicates that resistance to a given drug could not definitively be detected based on the test's algorithm.

***inhA* promoter**

Gene target included in the Xpert MTB/XDR test to detect MTB and resistance to isoniazid and ethionamide. Mutations in the *inhA* promoter region of TB are known to confer low level resistance to isoniazid and high-level cross resistance to ethionamide.

Intergenic region

Is a region of DNA sequence located between genes and a subset of noncoding DNA. Some intergenic regions act to control coding regions (genes) nearby.

katG

Gene target included in the Xpert MTB/XDR test to detect mutations that confer resistance to isoniazid.

Locus

A locus is the position of a genetic feature in the DNA sequence, like a genetic street address. Loci are standardized between genomes by reference to a common reference genome, such as H37Rv for *M tuberculosis*.

Microbiologically confirmed

Refers to a biological specimen that is positive by culture or a WHO-recommended rapid molecular test, such as Xpert MTB/RIF, Xpert Ultra, or Truenat MTB.

Mutation

A mutation is a change in a DNA sequence. Mutations can result from DNA copying mistakes made during cell division, exposure to ionizing radiation, exposure to chemicals called mutagens, or infection by viruses.

Non-determinate test result

A non-determinate Xpert MTB/XDR test result is one that results in an Error, Invalid, or No Result and can be due to an operator error, instrument, or cartridge issue.

***oxyR-ahpC* intergenic region**

Gene targets included in the Xpert MTB/XDR test to detect mutations that confer resistance to isoniazid.

Phenotypic drug susceptibility testing (pDST)

Phenotypic testing requires growth of *M tuberculosis* in the presence of antibiotics at a specific concentration that will inhibit the growth of a sensitive organism or have no impact on growth of a resistant organism.

Presumptive tuberculosis

Refers to a patient who presents with symptoms or signs suggestive of or compatible with tuberculosis.

Promoter region

A promoter region is a sequence of DNA where the transcriptional machinery binds before transcribing the DNA into RNA that may then be translated into an amino acid sequence.

Resistance-determining region

A region of the *M tuberculosis* genome where mutations commonly cause resistance to a specific drug.

rrs

Gene target included in the Xpert MTB/XDR test to detect mutations that confer resistance to second line injectable drugs, amikacin, kanamycin, and capreomycin.

Sanger sequencing

Technique for DNA sequencing based upon the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication, also known as 'the chain termination method'.

Targeted gene sequencing

The process for detecting predetermined mutations in DNA or genomic regions.

Whole genome sequencing (WGS)

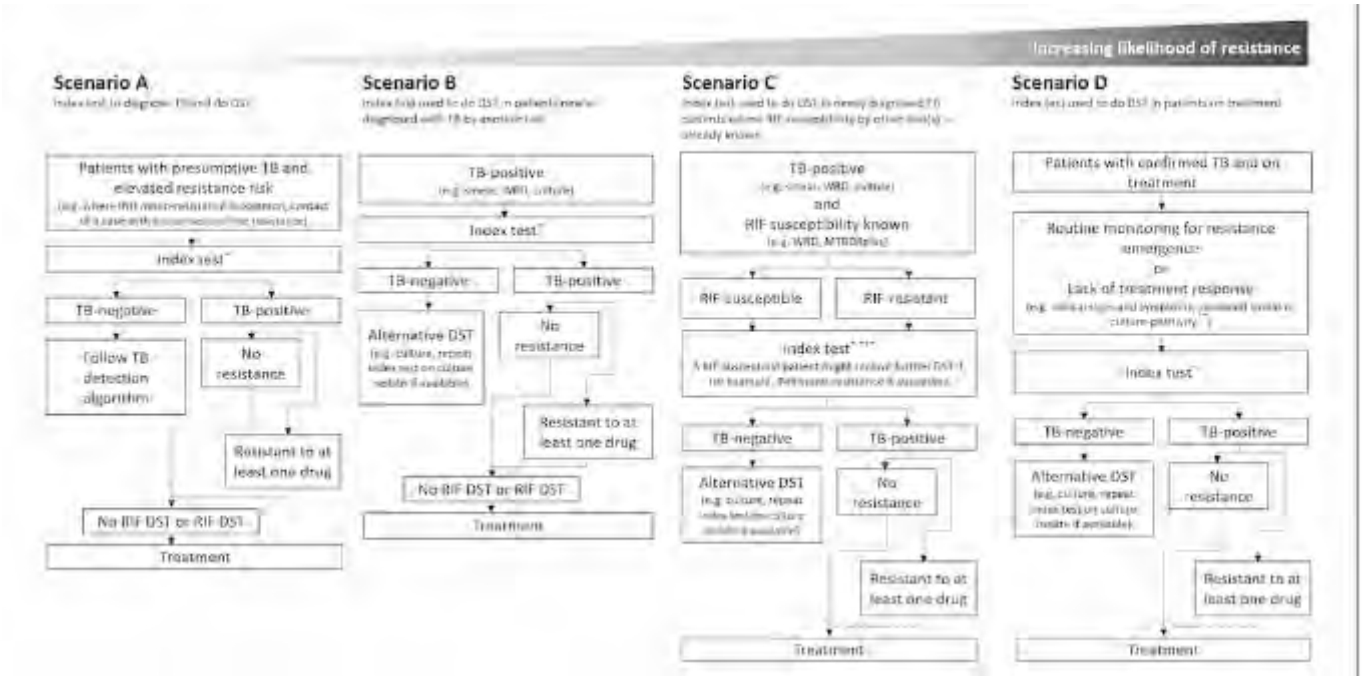
The process of determining the complete genome sequence for a given organism at one time through next generation sequencing methods. This method can determine the order of all nucleotides in a given genome and detect any variations relative to a reference genome using bioinformatics analyses.

Appendix 2. Possible test results for each target in the Xpert MTB/XDR assay

Drug Class	Result Call
	INVALID/ERROR/NO RESULT
N/A	MTB DETECTED
	MTB NOT DETECTED
Isoniazid	Low INH Resistance DETECTED
	INH Resistance DETECTED
	INH Resistance NOT DETECTED
	INH Resistance INDETERMINATE
Fluoroquinolone	Low FLQ Resistance DETECTED
	FLQ Resistance DETECTED
	FLQ Resistance NOT DETECTED
	FLQ Resistance INDETERMINATE
Amikacin	AMK Resistance DETECTED
	AMK Resistance NOT DETECTED
	AMK Resistance INDETERMINATE
Kanamycin	KAN Resistance DETECTED
	KAN Resistance NOT DETECTED
	KAN Resistance INDETERMINATE
Capreomycin	CAP Resistance DETECTED
	CAP Resistance NOT DETECTED
	CAP Resistance INDETERMINATE
Ethionamide*	ETH Resistance DETECTED
	ETH Resistance NOT DETECTED

Positive results for the Xpert MTB/XDR assay can be MTB DETECTED and all resistance targets are NOT DETECTED, or MTB DETECTED and one or more of the resistance targets is DETECTED, or MTB DETECTED and/or one or more of the following resistance targets is INDETERMINATE.

Appendix 3. Figure. Clinical pathway



The index test may be used in the following scenarios.
A. Index test used to diagnose tuberculosis and detect drug resistance.
B. Index test used to detect drug resistance in patients newly diagnosed with tuberculosis by another test where rifampicin susceptibility is unknown. Proposed role of Xpert MTB/XDR would be an initial test for resistance to isoniazid and second-line drugs (replacement for LPAs and culture-based DST as initial tests).
C. Index test used to detect drug resistance in patients newly diagnosed with tuberculosis and rifampicin resistance by other tests (although less likely, it is possible that the index test may still be done when documented rifampin susceptibility by other tests exists). Proposed role of Xpert MTB/XDR would be an initial test for resistance to isoniazid and second-line drugs (replacement for LPAs and culture-based DST as initial tests).
D. Index test used to detect drug resistance in patients on treatment. Proposed role of Xpert MTB/XDR would be a test used in combination with other tests for treatment monitoring (parallel testing).

Abbreviations: DST: drug susceptibility testing; RIF: rifampicin; TB: tuberculosis; WRD: WHO-recommended rapid diagnostic.

*Although direct testing is preferred for rapidity (which can be done on a raw specimen or a specimen remnant after some form of processing such as N-acetyl-L-cysteine (NALC)-NaOH decontamination), indirect testing using a cultured isolate could also be done (if, for example, a MTBC-positive reflex result is unavailable or culture has already been done due to diagnose tuberculosis).
**Xpert MTB/XDR may be considered in patients who were Xpert MTB/Ultra rifampicin susceptible prior to treatment and transitioned to Xpert MTB/Ultra rifampicin resistant while on treatment.
***Although index test use may be prioritised when risks of isoniazid- and/or second-line-resistance are elevated (in Scenario C if rifampicin resistance is first detected), it may also be applied irrespective of what the rifampicin susceptibility is, although we expect this to be less frequent.

Notes: 1) for all regimens, final composition will depend on other factors, including rifampicin susceptibility determined by an alternative test; 2) the timing of rifampicin DST can be before, in parallel, or after the index test is applied; and 3) for ease of presentation, tuberculosis and MTBC are treated equivalently.

Appendix 4. Search strategy

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R) <1946 to present>

---

1 Extensively Drug-Resistant Tuberculosis/ or Tuberculosis, Multidrug-Resistant/ or Tuberculosis, Pulmonary/

2 (tuberculosis adj3 (lung or pulmonary)).mp.

3 (tuberculosis adj3 respiratory).mp.

4 (isoniazid resistance or isoniazid resistant).mp.

5 ((Fluoroquinolone adj3 resistance) or (Fluoroquinolone adj3 resistant)).mp.

6 (Second-line injectable drug adj3 resistance).mp.

7 (Second-line injectable drug adj 3 resistant).mp.

8 ((SLID adj3 resistance) or (SLID adj3 resistant)).mp.

9 (MDR-TB or XDR-TB).mp.

10 ((isoniazid or fluoroquinolone or "second-line injectable drug" or SLID) adj3 monoresist*).mp.

11 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10

12 (cartridge adj3 test*).mp.

13 cartridge*.ab. or cartridge*.ti.

14 (Molbio or Truenat or Cepheid or Xpert* or Bioneer or Hain).mp.

15 Genexpert*.mp.

16 exp Point-of-Care Systems/

17 drug susceptibility test*.mp.

18 12 or 13 or 14 or 15 or 16 or 17

19 11 and 18

20 limit 19 to yr="2015 -Current"

Appendix 5. QUADAS-2

Domain 1: Patient selection

Detection of tuberculosis

Risk of bias: Could the selection of patients have introduced bias?

Signalling question 1: Was a consecutive or random sample of patients enrolled?

We answered yes if the study enrolled a consecutive or random sample of eligible patients; no if the study selected patients by convenience; and unclear if the study did not report the manner of patient selection or was not clearly reported.

Signalling question 2: Was a case-control design avoided?

We answered yes if the study enrolled patients with presumptive tuberculosis; no if the study enrolled cases with confirmed tuberculosis and controls from a healthy population; and unclear if we cannot tell. We consider that accuracy studies may have a cross-sectional design even when the reference standard is performed before the index test if both cases and controls are sampled from a single source population.

Signalling question 3: Did the study avoid inappropriate exclusions?

We answered yes if the study included both smear-positive and smear-negative individuals; no if the study included primarily or exclusively smear-positive or smear-negative patients; and unclear if we cannot tell. If at the time of specimen collection, the patient was on any form of tuberculosis treatment and if culture reference standard was used, we answered no because the bactericidal action of antibiotics can cause negative culture and positive PCR results.

Applicability: Are there concerns that the included patients and setting do not match the review question?

We answered low concern if patients were evaluated as outpatients (with either expectorated or induced sputum) in local hospitals or primary care centres. We answered high concern if patients were evaluated exclusively as inpatients in tertiary care centres. We answered unclear concern if the clinical setting was not reported or there was insufficient information to make a decision. We also answered unclear concern if testing was done at a central-level laboratory and the clinical setting was not reported if, for example, it was difficult to tell whether the laboratory provided services mainly to very sick patients or patients with a broader clinical spectrum of illness.

Detection of drug resistance

Risk of bias: could the selection of patients have introduced bias?

Signalling question 1: Was a consecutive or random sample of patients enrolled?

We answered the same as for detection of tuberculosis.

Signalling question 2: Was a case-control design avoided?

We answered yes if the study enrolled tuberculosis patients with suspected or sufficiently high pre-test probability (per WHO guidelines) for resistance to isoniazid, second-line drugs, or both isoniazid and second-line drugs; no if the study enrolled tuberculosis patients with confirmed pre-known resistance to the drug in question; and unclear for all other scenarios or if it was not clearly reported. We consider that accuracy studies may have a cross-sectional design even when the reference standard is performed before the index test if both cases and controls are sampled from a single source population.

Signalling question 3: did the study avoid inappropriate exclusions?

We answered yes for people who were previously treated for tuberculosis. we answered no if people who were previously treated were excluded. Patients previously tested for tuberculosis have a higher risk of having drug resistance and are likely to be the target population for initial use of the index tests. In people with samples known to be heteroresistant (a mix of susceptible and resistant tuberculosis strains in the specimen) were excluded, which is particularly relevant for fluoroquinolones, we answered no. We answered unclear if we cannot tell.

Applicability: Are there concerns that the included patients and setting do not match the review question?

We judged low concern if the selected clinical specimens or isolates match the review question, which reflects the way the test will be used in practice.

We judged high concern if the selected specimens or isolates did not represent those for whom the test will be used in practice, such as in individuals who do not require investigation for resistance to the drugs in question.

We will judge unclear concern if we cannot not tell.

Domain 2: Index test

Detection of tuberculosis

Risk of bias: could the conduct or interpretation of the index test have introduced bias?

Signalling question 1: were the index test results interpreted without knowledge of the results of the reference standard?
We answered this question yes for all studies where results are automatically generated and the user is provided with printable test results. Thus, there is no room for subjective interpretation of test results. For those assays, which require user interpretation, we answered yes if the reader of the assay was blinded to results of reference tests. We answered no if the reader of the assay was not blinded to the results of reference tests. If the specimens were from a biobank (repository that stores biological specimens) comprised of specimens with known second-line drug resistance and the identity of these specimens was known to the assay reader, we will also answer no unless the assay automatically generates results. We answered unclear if it was not stated in the paper or if the study authors failed to answer this question.
Signalling question 2: if a threshold was used, was it prespecified?
We answered yes for all studies.

Applicability: are there concerns that the index test, its conduct, or its interpretation differ from the review question?
Variations in test technology, execution, or interpretation may affect estimates of the diagnostic accuracy of a test. We will judge the study to be of low concern for applicability if the test was performed as recommended by the manufacturer. We judged the study to be of high concern if the test was applied differently than recommended by the manufacturer, for example if the test was applied to pooled sputa. We judged the study to be of unclear concern if we cannot tell.

Detection of drug resistance
Risk of bias: could the conduct or interpretation of the index test have introduced bias?

Signalling question 1. were the index test results interpreted without knowledge of the results of the reference standard?
We answered this question yes for all studies where results are automatically generated and the user is provided with printable test results, such as drug susceptibility testing run by MGIT 960 SIRE. For those assays which require user interpretation, such as Löwenstein–Jensen (LJ) drug susceptibility testing, we answered yes if the reader of the assay was blinded to results of reference tests. We answered no if the reader of the assay was not blinded to the results of reference tests. We answered unclear if it was not stated in the paper or if the study authors failed to answer this question.
Signalling question 2: if a threshold was used, was it prespecified?
We answered yes for all studies.

Applicability: are there concerns that the index test, its conduct, or its interpretation differ from the review question?
Same judgements as for detection of tuberculosis.

Domain 3: Reference standard
Detection of tuberculosis
Risk of bias: Could the reference standard, its conduct, or its interpretation have introduced bias?

Signalling question 1: Is the reference standard likely to correctly classify the target condition?
We answered yes for all studies, since a microbiological reference standard for *M tuberculosis* identification was a criterion for inclusion in the review.
Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?
We answered yes if the reference test provided an automated result (for example, MGIT 960), blinding was explicitly stated, or it was clear that the reference standard was performed at a separate laboratory and/or performed by different people. We answered no if the study stated that the reference standard result was interpreted with knowledge of the index test result. We answered unclear if we could not tell.

Applicability: Are there concerns that the target condition as defined by the reference standard does not match the question? We answered high concern if a type of culture was not done as part of the reference standard, because studies that include only DNA-based tests do not directly measure live M tuberculosis. We answered low concern if culture was performed. We answered unclear concern if we cannot tell.

Detection of drug resistance
Risk of bias: could the reference standard, its conduct or its interpretation have introduced bias?

Signalling question 1: is the reference standard likely to correctly classify the target condition?
We answered these questions for each target condition separately by reference standard as follows.

Drug	pDST	gDST, targeted sequencing	Composite (pDST and gDST, targeted sequencing)	gDST, whole genome sequencing)	Composite (pDST and gDST, whole genome sequencing)
Isoniazid	Yes	Unclear if few loci are investigated, and yes, if all relevant loci are analysed. Loci required for yes*: katG*, *inhA* promoter, *oxyR-ahpC i*ntergenic region, and *fabG1*	Yes	Unclear if few loci are investigated, and yes, if all relevant loci are analysed. Loci required for yes*: katG*, *inhA* promoter, *oxyR-ahpC i*ntergenic region, and *fabG1*	Yes
Fluoroquinolone	Yes, will depend on critical concentration used for moxifloxacin*	Yes Loci required for yes: *gyrA* and *gyrB*	Yes	Yes Loci required for yes: *gyrA* and *gyrB*	Yes
Ethionamide	No, there is considerable overlap in the MICs of *M tuberculosis* isolates with and without resistance-causing variants.	Unclear if few loci are investigated, and yes, if all relevant loci are analysed Loci required for yes: *ethA*, *ethR*, and i*nhA* promoter No if only the *inhA* promoter was analysed	Unclear	Unclear if few loci are investigated, and yes, if all relevant loci are analysed. Loci required for yes: e*thA*, *ethR*, and i*nhA* promoter No if only the *inhA* promoter was analysed	Unclear
Amikacin	Yes	Yes, if all relevant loci are analysed Loci required for yes: *rrs* and *eis* promoter	Yes	Yes, if all relevant loci are analysed Loci required for yes: *rrs* and *eis* promoter	Yes

Abbreviations: gDST: genotypic drug susceptibility testing; pDST: phenotypic drug susceptibility testing.

*We used the currently-recommended WHO critical concentrations as a benchmark for judging risk of bias. For *M tuberculosis*, the antimicrobial susceptibility testing critical concentration is defined as the lowest concentration of an anti-tuberculosis agent in vitro that will inhibit the growth of 99% of phenotypically wild type strains of *M tuberculosis* complex", (WHO Critical concentrations 2018).

Signalling question 2: were the reference standard results interpreted without knowledge of the results of index test.
For pDST, we answered yes if the reference test provided an automated result (for example, if liquid culture is used as in MGIT 960 DST), blinding was explicitly stated, or it was clear that the reference test was performed at a separate laboratory, or performed by different people, or both. Of note, pDST on solid media is not automated. We answered no if the study stated that the reference standard result was interpreted with knowledge of the index test result. We answered unclear if we cannot tell. For gDST, we answered yes for all studies since the results for the reference standard are automated.

We added the following signalling question.
Signalling question 3: Were the index test and reference standard both done on material of the same type (clinical specimen or sediment, or isolate)?
Phenotypic DST (pDST) and genotypic DST (gDST) for reference standard testing can be done on an isolate that has undergone (potentially multiple rounds) of culture in drug-free media. This may lead to the depletion of

resistant strains present in the original specimen (which would have been used for the index test if direct testing was done) and cause discrepant results. We think this is an important question as it addresses heteroresistance, which often explains discordance between genotypic and phenotypic results.
For direct testing of a clinical specimen by the index test: we answered yes if the reference test was done directly on the same clinical specimen; no if the reference standard was done on a culture isolate; and unclear if we could not tell. For indirect testing of a culture isolate by the index test: we answered yes if the reference test was done on the same culture isolate (e.g. indirect sequencing); no if the reference standard was done on a different culture isolate, or specimen; and unclear if we could not tell.

Applicability: are there concerns that the target condition as defined by the reference standard does not match the question?
We judged applicability to be of low concern for all studies because specimens to be subsequently tested for drug resistance will have already been identified as *M tuberculosis* complex positive.

Domain 4: Flow and timing
Detection of tuberculosis
Risk of bias: could the patient flow have introduced bias?

Signalling question 1: was there an appropriate interval between the index test and reference standard?
We expect the reference standard test to be undertaken at the same time as the index test (i.e. each performed on a paired sample for most studies). However, we expected some studies to include specimens from patients who had received a reference test on an earlier sample. The sample applies to some culture isolates, whose drug susceptibility profile might have been confirmed prior to the index test being available. We answered yes if the tests were paired or were separated by a few days. We answered no if reference and index tests were not done on paired samples and were separated by several months. As patients suspected of second-line drug resistance are often on some form of anti-tuberculosis therapy, it is possible that variation in the microbial population of specimens collected at different time points may occur. We answered unclear if it was not stated in the paper or if the authors failed to answer this question.
Signalling question 2: did all patients receive the same reference standard?
We answered yes if the reference standard was applied to all patients or a random sample of patients, no if the reference standard was only applied to a selective group of patients, and unclear if it was not stated in the paper or if the authors failed to answer this question.
Signalling question 3: were all patients included in the analysis?
We determined the answer to this question by comparing the number of participants enrolled with the number of patients included in the 2 x 2 tables. We will note if the study authors reported the number of indeterminate assay results. We answered yes if the number of participants enrolled was clearly stated and corresponded to the number presented in the analysis or if exclusions were adequately described. We answered no if there were participants missing or excluded from the analysis and there was no explanation given. We answered unclear if not enough information was given to assess whether participants were excluded from the analysis.

Detection of drug resistance
We answered the same as for detection of tuberculosis.

Judgements for risk of bias assessments for a given domain
If we answered all signalling questions for a domain yes, then we judged risk of bias as low.
If we answered all or most signalling questions for a domain no, then we judged risk of bias as high.
If we answered only one signalling question for a domain no, we discussed further the risk of bias judgement.
If we answered all or most signalling questions for a domain unclear, then we judged risk of bias as unclear.
If we answered only one signalling question for a domain unclear, we discussed further the risk of bias judgement for the domain.
For reference standard domain, if either any of the reference standard had signalling no, we judged risk of bias as high.

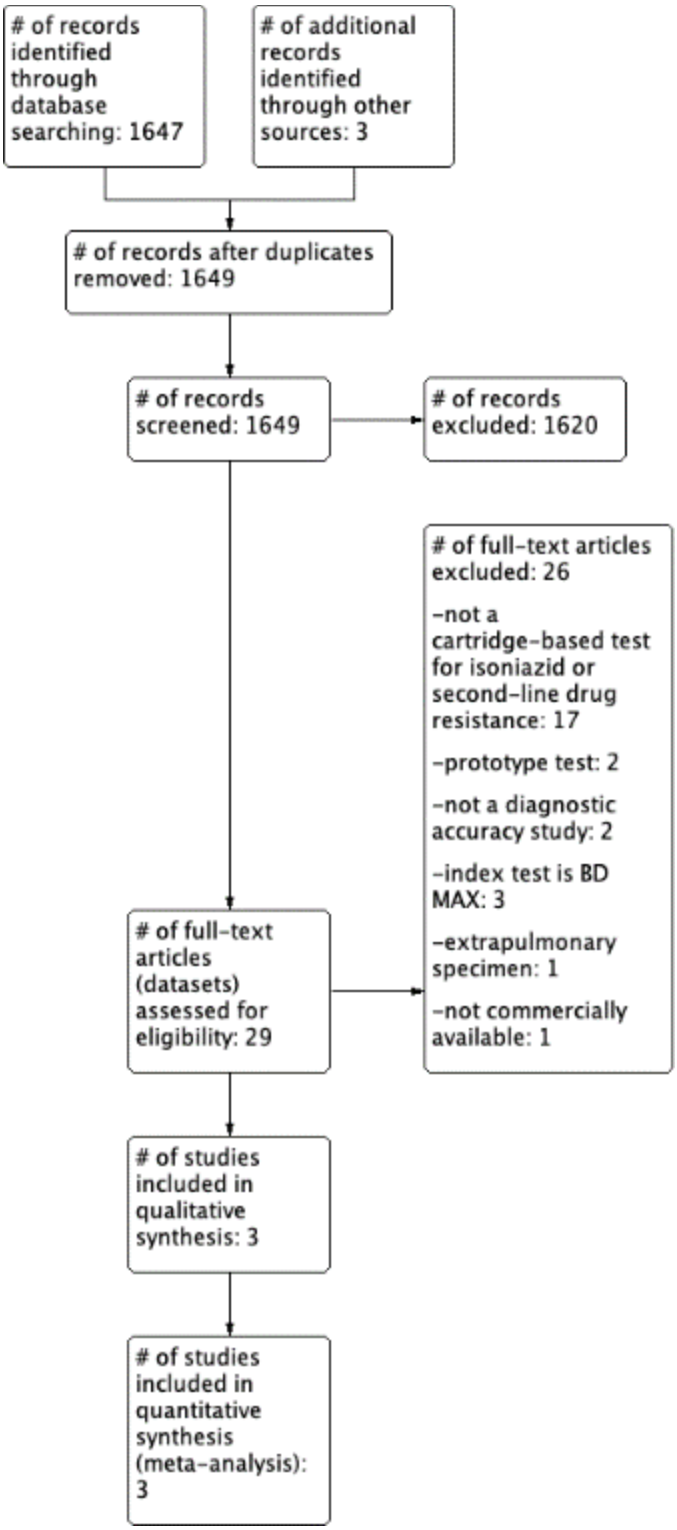
Critical concentrations and clinical breakpoints for medicines recommended for the treatment of rifampicin-resistant and multidrug-resistant tuberculosis

Drug groups	Drug	LJ	7H10	7H11	MGIT
First-line	Isoniazid	0.2	0.2	0.2	0.1
Fluroquinolones	Levofloxacin	2.0	1.0	-	1.0
	Moxifloxacin (CC)	1.0	0.5	0.5	0.25
	Moxifloxacin (CB)	–	2.0	-	1.0

	Gatifloxacin	0.5	-	-	0.25
Second-line	Amikacin	30.0	2.0	-	1.0
Other	Ethionamide	40.0	5.0	10.0	5.0

Abbreviations: CB: critical breakpoint; CC: critical concentration

Appendix 6. Flow of studies in the review



Appendix 7. List of studies

Included studies

Cepheid 2020
Clinical evaluation of the Xpert® MTBXDR assay
Report R244C2 Xpert MTB/XDR Rev 1.0
Sponsor: Cepheid, Sunnyvale, USA

DIAMA 2020
DIAgnostics for Multidrug Resistant Tuberculosis in Africa
ClinicalTrials.gov Identifier: NCT03303963
Sponsor: Dissou Affolabi, Kigali, Rwanda

FIND 2020
Analytical Performance and Clinical Diagnostic Accuracy of the Xpert MTB/XDR Assay for TB and Expanded Resistance Detection, September 2020
ClinicalTrials.gov Identifier: NCT03728725
Sponsor: Foundation for Innovative New Diagnostics, Geneva, Switzerland

Excluded studies

1. Andreevskaya SN, Smirnova TG, Larionova EE, Andrievskaya IY, Chernousova LN, Ergeshov A. Isoniazid-resistant Mycobacterium tuberculosis: prevalence, resistance spectrum and genetic determinants of resistance. Bulletin of Russian State Medical University. 2020 (1):21-6. (Not a cartridge-based test for isoniazid or second-line drug resistance)
2. Beutler M, Plesnik S, Mihalic M, Olbrich L, Heinrich N, Schumacher S, et al. A pre-clinical validation plan to evaluate analytical sensitivities of molecular diagnostics such as BD MAX MDR-TB, Xpert MTB/Rif Ultra and FluoroType MTB. PLOS One. 2020;15(1):e0227215. (Not a diagnostic accuracy study)
3. Bisognin F, Lombardi G, Finelli C, Re MC, Dal Monte P. Simultaneous detection of Mycobacterium tuberculosis complex and resistance to rifampicin and isoniazid by MDR/MTB ELITe MGB R kit for the diagnosis of tuberculosis. PLOS One. 2020;15(5):e0232632. (Not a cartridge-based test for isoniazid or second-line drug resistance)
4. Broda A, Nikolayevskyy V, Casali N, Khan H, Bowker R, Blackwell G, et al. Experimental platform utilising melting curve technology for detection of mutations in Mycobacterium tuberculosis isolates. European Journal of Clinical Microbiology & Infectious Diseases. 2018;37(7):1273-9. (Not a cartridge-based test for isoniazid or second-line drug resistance)
5. Chakravorty S, Roh SS, Glass J, Smith LE, Simmons AM, Lund K, et al. Detection of isoniazid-, fluoroquinolone-, amikacin-, and kanamycin-resistant tuberculosis in an automated, multiplexed 10-Color assay suitable for point-of-care use. Journal of Clinical Microbiology. 2017;55(1):183-98. (Prototype test)
6. Chang Y, Kim S, Kim Y, Ei PW, Hwang D, Lee J, et al. Evaluation of the QuantaMatrix multiplexed assay platform for molecular diagnosis of multidrug- and extensively drug-resistant tuberculosis using clinical strains isolated in Myanmar. Annals of Laboratory Medicine. 2020;40(2):142-7. (Not a cartridge-based test for isoniazid or second-line drug resistance)
7. Chumpa N, Kawkitinarong K, Rotcheewaphan S, Sawatpanich A, Petsong S, Tumwasorn S, et al. Evaluation of Anyplex TM II MTB/MDR kit's performance to rapidly detect isoniazid and rifampicin resistant Mycobacterium tuberculosis from various clinical specimens. Molecular Biology Reports. 2020;47(4):2501-8. (Not a cartridge-based test for isoniazid or second-line drug resistance)
8. Ciesielczuk H, Kouvas N, North N, Buchanan R, Tiberi S. Evaluation of the BD MAX TM MDR-TB assay in a real-world setting for the diagnosis of pulmonary and extra-pulmonary TB. European Journal of Clinical Microbiology & Infectious Diseases.2020;39(7):1321-7. (Index test is BD MAX)
9. Foongladda S, Banu S, Pholwat S, Gratz J, O-Thong S, Nakkerd N, et al. Comparison of TaqMan(R) Array Card and MYCOTB(TM) with conventional phenotypic susceptibility testing in MDR-TB. International Journal of Tuberculosis and Lung Disease. 2016;20(8):1105-12. (Not a cartridge-based test for isoniazid or second-line drug resistance)
10. Galarza M, Fasabi M, Levano KS, Castillo E, Barreda N, Rodriguez M, et al. High-resolution melting analysis for molecular detection of multidrug resistance tuberculosis in Peruvian isolates. BMC Infectious Diseases. 2016;16:260. (Not a cartridge-based test for isoniazid or second-line drug resistance)
11. Han Y, Xiao N, Huang S, Qin M, Che N, Liu Z. The Application of Xpert Mycobacterium tuberculosis/rifampicin, quantitative polymerase chain reaction and high resolution melting curve in the diagnosis of superficial lymph node TB. Current Pharmaceutical Biotechnology. 2019;20(12):1044-54. (Extrapulmonary specimens)
12. Havlicek J, Dachsel B, Slickers P, Andres S, Beckert P, Feuerriegel S, et al. Rapid microarray-based detection of rifampin, isoniazid, and fluoroquinolone resistance in Mycobacterium tuberculosis by use of a single cartridge. Journal of Clinical Microbiology. 2018;56(2). (Not a cartridge-based test for isoniazid or second-line drug resistance)
13. Huang F, Dang L, Sun H, Yang H, Wu X. [A study of the value of three molecular diagnostic techniques in the diagnosis of tuberculosis]. Zhonghua jie he he hu xi za zhi = Zhonghua jiehe he huxi zazhi = Chinese Journal of

Tuberculosis and Respiratory Diseases. 2015;38(9):680-5. (Not a cartridge-based test for isoniazid or second-line drug resistance)
14. Kim S, Kim Y, Chang Y, Hirgo WK, Chang CL, Shim T-S, et al. Comparison of Quantamatrix multiplexed assay platform and GenoType MTBDR assay using smear-positive sputum specimens from patients with multidrug-resistant/extensively drug-resistant tuberculosis in South Korea. Frontiers in Microbiology. 2019;10:1075. (Not a cartridge-based test for ISONIAZID or second-line drug resistance)
15. Klotoe BJ, Molina-Moya B, Gomes HM, Gomgnimbou MK, Oliveira Suzarte L, Feres Saad MH, et al. TB-EFI, a novel 18-Plex microbead-based method for prediction of second-line drugs and ethambutol resistance in Mycobacterium tuberculosis complex. Journal of Microbiological Methods. 2018;152:10-7. (Not a cartridge-based test for isoniazid or second-line drug resistance)
16. Law ILG, Loo JFC, Kwok HC, Yeung HY, Leung CCH, Hui M, et al. Automated real-time detection of drug-resistant Mycobacterium tuberculosis on a lab-on-a-disc by recombinase polymerase amplification. Analytical Biochemistry. 2018;544:98-107. (Not a cartridge-based test for isoniazid or second-line drug resistance)
17. Lee YS, Kang MR, Jung H, Choi SB, Jo K-W, Shim TS. Performance of REBA MTB-XDR to detect extensively drug-resistant tuberculosis in an intermediate-burden country. Journal of Infection and Chemotherapy. 2015;21(5):346-51. (Not a cartridge-based test for isoniazid or second-line drug resistance)
18. Li Q, Ou XC, Pang Y, Xia H, Huang HR, Zhao B, et al. A novel automatic molecular test for detection of multidrug resistance tuberculosis in sputum specimen: A case control study. Tuberculosis (Edinburgh, Scotland). 2017;105:9-12. (Not commercially available)
19. Mokaddas EM, Ahmad S, Eldeen HS. GeneXpert MTB/RIF is superior to BBD Max MDR-TB for diagnosis of tuberculosis (TB) in a country with low incidence of multidrug-resistant TB (MDR-TB). Journal of Clinical Microbiology. 2019;57(6). (Index test is BD MAX)
20. Murray P, Cooper C, Maus C. Comparative Performance of BD MAX MDR-TB and Cepheid Xpert MTB/RIF Assays. Journal of Clinical Microbiology. 2019;57(9). (Not a diagnostic accuracy study)
21. Pang Y, Dong H, Tan Y, Deng Y, Cai X, Jing H, et al. Rapid diagnosis of MDR and XDR tuberculosis with the MeltPro TB assay in China. Scientific Reports. 2016;6:25330. (Not a cartridge-based test for isoniazid or second-line drug resistance)
22. Santos PFGD, Costa ERD, Ramalho DM, Rossetti ML, Barcellos RB, Nunes LdS, et al. Detection of tuberculosis drug resistance: a comparison by Mycobacterium tuberculosis MLPA assay versus Genotype RMTBDRplus. Memorias do Instituto Oswaldo Cruz. 2017;112(6):396-403. (Not a cartridge-based test for isoniazid or second-line drug resistance)
23. Shah M, Paradis S, Betz J, Beylis N, Bharadwaj R, Caceres T, et al. Multicenter study of the accuracy of the BD MAX TM MDR-TB assay for detection of Mycobacterium tuberculosis complex and mutations associated with resistance to rifampin and isoniazid. Clinical Infectious Diseases. 2019. (Index test is BD MAX)
24. Strydom K, Ismail F, Matabane MMZ, Onwuegbuna O, Omar SV, Ismail N. Comparison of three commercial molecular assays for detection of rifampin and isoniazid resistance among Mycobacterium tuberculosis isolates in a High-HIV-prevalence setting. Journal of Clinical Microbiology. 2015;53(9):3032-4. (Not a cartridge-based test for isoniazid or second-line drug resistance)
25. Wang HY, Uh Y, Kim S, Cho E, Lee JS, Lee H. Detection of rifampicin- and isoniazid-resistant Mycobacterium tuberculosis using the Quantamatrix multiplexed assay platform system. Annals of Laboratory Medicine. 2018;38(6):569-77. (Not a cartridge-based test for ISONIAZID or second-line drug resistance)
26. Xie YL, Chakravorty S, Armstrong DT, Hall SL, Via LE, Song T, et al. Evaluation of a Rapid Molecular Drug-Susceptibility Test for Tuberculosis. The New England journal of medicine. 2017;377(11):1043-54. (Prototype test)

Appendix 8. Table. Characteristics of included studies

Study year	Countries/centres (High MDR Burden)	Study design	Number of patients (% detected RR)	Age of enrolment	PLHIV	Reference standard for drug resistance	Loci included in gDST reference standard
Cepheid 2020	China (yes) South Africa (yes)	cross-sectional, retrospective[1]	530 (47.9%)	≥ 15 years[2]	NR	pDST gDST composite	*katG, inhA* promoter, *oxyR-ahpC* intergenic region, *fabG1, gyrA, gyrB, rrs*, *eis* promoter
DIAMA 2020	Benin (no) Cameroon (no) Rwanda (no)	cross-sectional, prospective	621 (48.3%)	≥ 15 years	13.3% Benin; 37.2% Rwanda	pDST	NA
FIND 2020	New Delhi (yes) Moldova (yes) Mumbai (yes) South Africa (yes)	cross-sectional, prospective	611 (80.9%)	≥ 18 years	17.5% overall, 87.1% South Africa	pDST gDST composite	*katG, inhA* promoter, *oxyR-ahpC* intergenic region, *fabG1, gyrA, gyrB, rrs, eis* promoter

Abbreviations: NA: not applicable; NR: not reported; RR: rifampicin resistance

Footnotes

[1]. In some cases, the reference standard was performed before and in other cases after the index test.

[2]. One participant was 13 years old.

SUPPLEMENTARY INFORMATION

Supplement A. MeltPro® XDR-TB

Background
On 30 October 2020, we were notified by the WHO Global Tuberculosis Programme about a clinical study conducted in China evaluating MeltPro® XDR-TB (Xiamen Zeesan Biotech Co., Ltd., China), a low complexity test for resistance to anti-tuberculosis drugs. The WHO provided us with a report summarizing the clinical study. We corresponded with study authors for additional information and clarifications.

MeltPro XDR-TB is a commercially available, low complexity test for detection of mutations associated with resistance to rifampicin, isoniazid, fluoroquinolones, and injectable second-line drugs. MeltPro XDR-TB is designed to detect drug resistance on specimens determined to be TB positive. MeltPro XDR-TB testing is performed using an all-in-one machine, Sanity 2.0, Figure 1. Manual pipetting is required in the preliminary sample preparation stage, and subsequent processes - nucleic acid extraction, sample loading, detection (i.e. real-time PCR), and interpretation of results - are all fully automatic. The detection of drug resistance is based on multi-color melting curve analysis.

Figure S1. Sanity 2.0

Regarding rifampicin, MeltPro XDR-TB bases detection of resistance on mutations in the *rpoB* gene.
Regarding isoniazid, MeltPro XDR-TB bases detection of resistance on mutations in the *ahpC* promoter region, *inhA* promoter region, and *katG gene.*
Regarding fluoroquinolones, MeltPro XDR-TB bases detection of resistance on mutations in the *gyrA* quinolone resistance determining region.
Regarding second-line injectable drugs, MeltPro XDR-TB bases detection of resistance on mutations in *rrs* gene and *eis* promoter region.

Search methods for identification of studies
On 4 November 2020, we ran an additional electronic search using the terms Zeesan and MeltPro. We did not identify any publications. In correspondence, the authors wrote, "We have not published relevant research reports yet. We expect to entrust the hospital with further clinical verification in the near future, and then publish relevant articles, "(personal communication, Lili Zheng, Xiamen Zeesan Biotech Co., Ltd, llzheng@zsandx.com, 10 November 2020).

Summary of the clinical study
This was a cross-sectional, prospective study in which participants were selected by convenience. The study aim was to evaluate and validate the performance of the MeltPro® XDR-TB test kit. The study authors did not collect information on participant characteristics, such as age, HIV status, and tuberculosis treatment history. Participants came from both outpatient and inpatient settings. The study was conducted in China, a high TB burden, high TB/HIV burden country, and high MDR-TB burden country.

Participants

Sputum samples were selected from patients who had submitted specimens for culture and subsequent drug-susceptibility testing during the routine work of the clinical laboratory of the hospital. All selected samples were first tested by the MeltPro®MTB Test Kit (Xiamen Zeesan Biotech Co., Ltd.) and if found to be TB positive, the sample was eligible for inclusion. The authors stated, "In other words, samples selected for enrolment were by convenience." Samples were included if they were tuberculosis positive, had DST results, and at least 2 mL were available after the laboratory had completed other tests.

Sample size = 715
- patients with presumptive tuberculosis, n = 592, outpatient setting
- patients suspected of having XDR-TB or pre-XDR-TB, n = 123, drug-resistant ward, inpatient setting

Index test was MeltPro XDR-TB

Reference standard was MGIT DST

Outcomes were sensitivity and specificity

Indeterminate results were not included in the determination of sensitivity and specificity.

Sequencing was performed to resolve discordant index test and culture-based DST results.

Results

Methodological quality assessment
In the patient selection domain, we judged this study to be of high risk of bias because participants were selected by convenience. Regarding applicability, we rated this as unclear because demographic information was not reported. For the other QUADAS-2 domains, index test, reference standard, and flow and timing, we judged low risk of bias. Regarding applicability, we considered the index test and reference standard domains to be of low concern.

Findings

See Figure S2.

- MeltPro sensitivity was 85% for resistance to isoniazid in people with rifampicin susceptibility and 88% in people with rifampicin resistance.
- MeltPro sensitivity was 90% for resistance to fluoroquinolones in people with rifampicin susceptibility and 91% in people with rifampicin resistance.
- MeltPro sensitivity was 88% for resistance to fluoroquinolones in people with rifampicin susceptibility and 79% in people with rifampicin resistance.
- MeltPro specificity was ≥ 97% for all drugs in people with rifampicin susceptibility, but lower (86% to 90%) in people with rifampicin resistance.
- Inconclusive results: there were 27/715 (3.8%), 27/715 (3.8%) and 19/715 (3.4%) clinical sputum specimens without valid signals for isoniazid, fluoroquinolones, and amikacin, respectively, which the authors thought could be caused by low concentrations of tuberculosis bacteria.

Sequencing to resolve discordant results

Isoniazid: There were 18 samples whose DST was isoniazid sensitive but detected as isoniazid resistant by MeltPro XDR-TB. Sequencing results showed that all of them had mutations in the detection region of probes. There were 22 samples whose DST was isoniazid resistant but detected as

isoniazid sensitive by MeltPro XDR-TB. Sequencing results showed that none of them showed any mutation in the coverage area of the probes.
Fluoroquinolones: There were 20 samples whose DST was fluoroquinolone sensitive but detected as fluoroquinolone resistant by MeltPro XDR-TB. Sequencing results showed that all of them had mutations in *gyrA*. There were 10 samples whose DST was fluoroquinolone resistant but detected as fluoroquinolone sensitive by MeltPro XDR-TB. Sequencing results showed that two samples had D94G heterogenic mutation, while the remaining eight samples showed no mutation in the coverage area of the probe.
Amikacin: There were 20 samples whose DST was amikacin sensitive but detected as amikacin resistant by MeltPro XDR-TB. Sequencing results showed that all of them had mutations in *rrs* gene. There were 10 samples whose DST was amikacin resistant but detected as amikacin sensitive by MeltPro XDR-TB. Sequencing results showed that none of them showed any mutation in the coverage area of the probes.

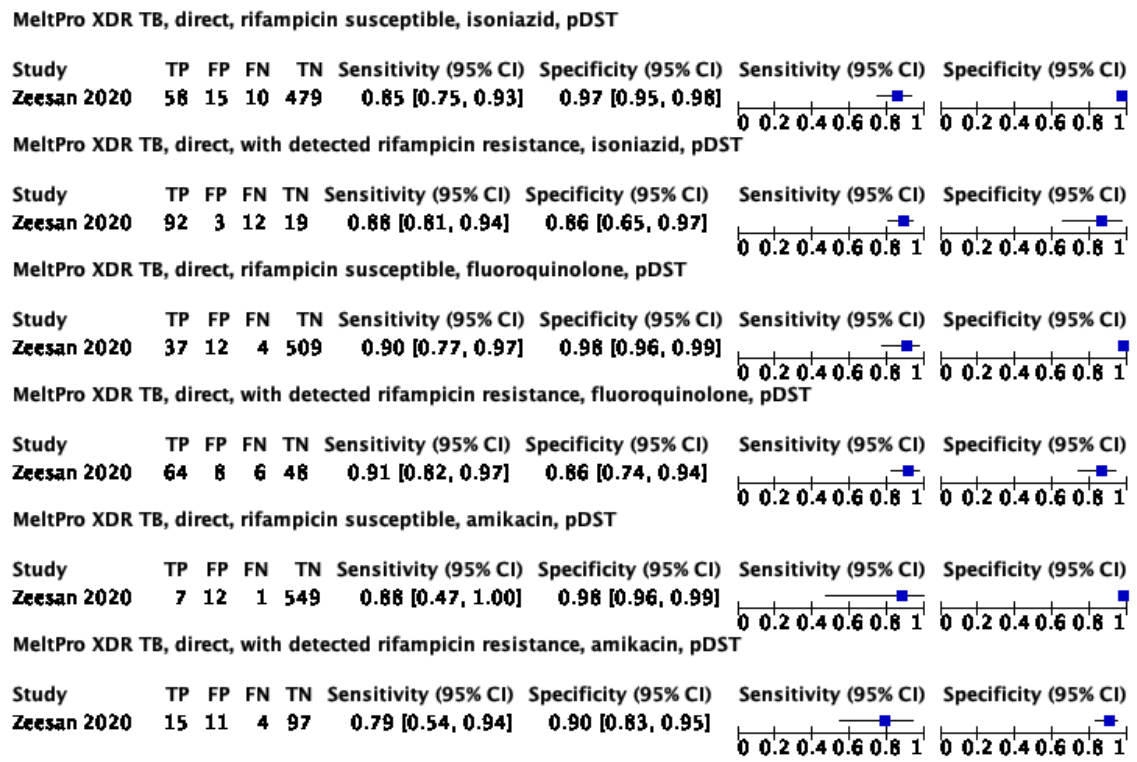


Figure S2. Forest plots of MeltPro XDR-TB sensitivity and specificity for detection of resistance to isoniazid, fluoroquinolones (levofloxacin), and amikacin by rifampicin resistance status.

Supplement B. Should Xpert MTB/XDR assay on sputum be used to diagnose PTB in people with signs and symptoms of TB?

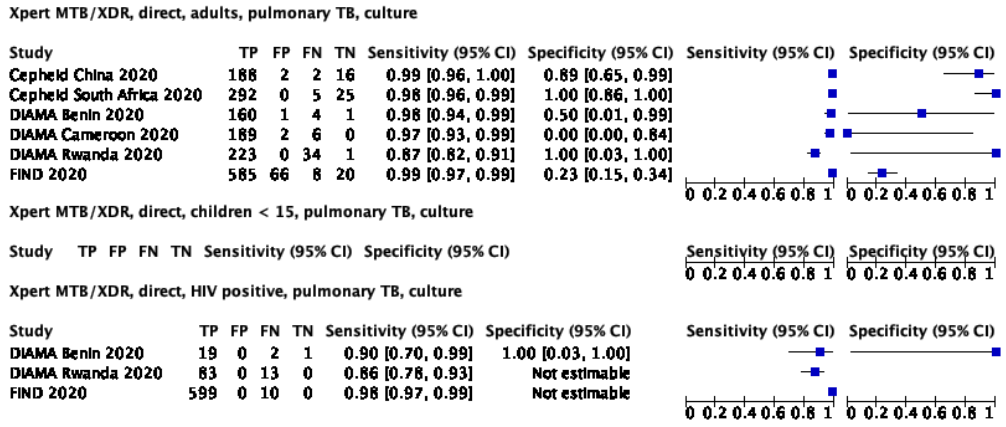


Figure. Forest plots of Xpert MTB/XDR sensitivity and specificity for the diagnosis of pulmonary tuberculosis by population, culture reference standard. Direct refers to testing directly on sputum

Supplement C. GRADE evidence profiles

1. Question: Should MTB/XDR assay on sputum be used to diagnose INH resistance in patients with microbiologically confirmed pulmonary TB, irrespective of resistance to RIF, pDST?

Sensitivity	0.94 (95% CI: 0.89 to 0.97)
Specificity	0.98 (95% CI: 0.95 to 0.99)

Prevalences	2%	10%	15%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of 2%	pre-test probability of 10%	pre-test probability of 15%	
True positives (patients with INH resistance)	3 studies 994 patients	cross-sectional (cohort type accuracy study)	not serious	serious [a]	not serious [b]	not serious	none	19 (18 to 19)	94 (89 to 97)	141 (134 to 146)	⊕⊕⊕◯ MODERATE
False negatives (patients incorrectly classified as not having INH resistance)								1 (1 to 2)	6 (3 to 11)	9 (4 to 16)	
True negatives (patients without INH resistance)	3 studies 611 patients	cross-sectional (cohort type accuracy study)	not serious	serious [a]	not serious	not serious	none	960 (933 to 972)	882 (857 to 893)	833 (809 to 843)	⊕⊕⊕◯ MODERATE
False positives (patients incorrectly classified as having INH resistance)								20 (8 to 47)	18 (7 to 43)	17 (7 to 41)	

Explanations

a. We had several concerns about whether there is indirectness in the populations studied. First, the median prevalence of isoniazid resistance in the included studies was 67.2% (range, 26.8% (DIAMA, Benin) to 93.9% (FIND, Moldova), higher than the three prevalences in the GRADE table. Applicability to settings with a lower prevalence of isoniazid resistance comes with some uncertainty. Second, there are potential differences in the mutations present in isoniazid mono-resistant strains and MDR strains. That is, there are studies that suggest that a more diverse set of mutations can be found in mono-resistant strains that MDR strains. Third, although the population for this PICO question is 'irrespective of rifampicin resistance,' owing to enrollment criteria in the studies, we note that most participants were rifampicin resistant. We downgraded one level for indirectness.

b. Sensitivity estimates ranged from 81% (FIND, New Delhi) to 100% (DIAMA, Rwanda). Regarding the low sensitivity estimate in New Delhi, the study authors reported that sequencing did not show the presence of variants typically associated with resistance in many phenotypically isoniazid-resistant samples suggesting that variants not analyzed by Xpert MTB/XDR might play a role. We did not downgrade for inconsistency. This was a judgement.

Web Annex 4.16. Drug concentrations used in culture-based drug susceptibility testing for included studies on second-line line probe assays diagnostic accuracy[7]

Table 1. Ofloxacin, levofloxacin, and moxifloxacin, drug concentrations used in culture-based drug susceptibility testing in relation to the WHO-recommended critical concentrations

Study	Reference standard	Concentration used (ug/ml)	Met WHO-recommended critical concentration	Comments
Ajbani 2012	MGIT 960	Ofloxacin: 2.0	Yes	
		Moxifloxacin: 0.25	No	
Barnard 2012	Middlebrook 7H11 (agar proportion)	Ofloxacin: 2.0	Yes	
Brossier 2010	LJ (agar proportion)	Ofloxacin: 2.0	No	
Catanzaro 2015	MGIT 960	Ofloxacin: 2.0	Yes	
		Moxifloxacin: 0.25	No	
Chikamatsu 2012	Ogawa	Levofloxacin: 1.0	Not applicable	No WHO- recommended concentration specified for Ogawa media
Fan 2011	MGIT 960	Ofloxacin: 2.0	Yes	
		Moxifloxacin: 0.25	No	
Ferro 2013	Middlebrook 7H10 (agar proportion)	Moxifloxacin: 2.0	Yes	
FIND 2016	MGIT 960	Ofloxacin: 2.0	Yes	
		Levofloxacin: 1.5	Yes	
		Moxifloxacin: 0.5	No	The WHO- recommended low level concentration was used

[7] Prepared by Christopher Gilpin and Alexei Korobitsyn (Both WHO)

Hillemann 2009	MGIT 960 and LJ	Ofloxacin: 2.0 for both media	Yes for MGIT; no for LJ	
Huang 2011	MGIT 960 and Middlebrook 7H11	Ofloxacin: 2.0	Yes for both media	
Ignatyeva 2012	MGIT 960	Ofloxacin: 2.0	Yes	
Jin 2013	LJ and BacT/ALERT 3D	Ofloxacin: 5.0 (LJ); 50 (BacT/ALERT 3D)	No for LJ; Not applicable BacT/ALERT 3D	No WHO- recommended concentration specified for BacT/ALERT 3D media
Kambli 2015a	MGIT 960	Ofloxacin: 2.0	Yes	
		Moxifloxacin: 0.25	No	
Kambli 2015b	MGIT 960	Levofloxacin: 1.5	Yes	
Kiet 2010	LJ	Ofloxacin: 2.0	No	
Kontsevaya 2011	MGIT 960	Ofloxacin: 2.0	Yes	
		Moxifloxacin: 0.25	No	
Kontsevaya 2013	MGIT 960	Ofloxacin: 2.0	Yes	
		Moxifloxacin: 0.25	No	
Lacoma 2012	BACTEC 460TB	Moxifloxacin: 0.5	Not applicable	No WHO- recommended concentration specified for BACTEC 460 media
Lopez-Roa 2012	MGIT 960 and Middlebrook 7H11	Ofloxacin: 2.0	Yes	
Miotto 2012	MGIT 960	Ofloxacin: 2.0	Yes	
NICD 2015	MGIT 960	Ofloxacin: 2.0	Yes	
Said 2012	Middlebrook 7H11	Ofloxacin: 2.0	Yes	
Simons 2015	MGIT 960 and Middlebrook 7H10 (agar dilution)	Moxifloxacin: 0.5 for MGIT and 1.0 for 7H10	No	For moxifloxacin using MGIT, the WHO-recommended low level concentration was used
Tagliani 2015	MGIT 960 and LJ (agar proportion)	Ofloxacin: 2.0 for MGIT and 4.0 for LJ	Yes	
	MGIT 960	Moxifloxacin: 0.5	No	The WHO- recommended low level concentration was used
	MGIT 960	Levofloxacin: 1.5	Yes	

Tomasicchio 2016	MGIT 960	Ofloxacin: 2.0	Yes	
Tukvadze 2014	LJ (proportion method)	Ofloxacin: 2.0	No	
van Ingen 2010	Middlebrook 7H10 (agar proportion)	Moxifloxacin: 1.0	No	
Zivanovic 2012	MGIT 960 and LJ agar proportion	Ofloxacin: 2.0 for both media	Yes for MGIT; no for LJ	

LJ-Löwenstein-Jensen; MGIT –Mycobacterial growth indicator tube

Reference: WHO, Updated interim critical concentrations for first-line and second-line DST (as of May 2012)
http://www.stoptb.org/wg/gli/assets/documents/Updated%20critical%20concentration%20table_1st%20and%202nd%20line%20drugs.pdf

Table 2. Amikacin, kanamycin, and capreomycin, drug concentrations used in culture-based drug susceptibility testing in relation to the WHO-recommended critical concentrations

Study	Reference standard	Concentration used (ug/ml)	Met WHO-recommended critical concentration	Comments
Ajbani 2012	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 2.5	Yes	
		Capreomycin: 2.5	Yes	
Barnard 2012	Middlebrook 7H11 (agar proportion)	Amikacin: 4.0	Not applicable	No WHO- recommended concentration specified for amikacin using 7H11
Brossier 2010	LJ (agar proportion)	Amikacin: 20.0	No	
		Kanamycin: 20.0	No	
		Capreomycin: 20.0	No	
Catanzaro 2015	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 2.5	Yes	
		Capreomycin: 2.5	Yes	
Chikamatsu 2012	Ogawa	Amikacin: unknown	Not applicable	No WHO- recommended concentration specified for Ogawa media
		Kanamycin: unknown	Not applicable	

		Capreomycin: unknown	Not applicable	
Fan 2011	MGIT 960	Amikacin: 1.0	Yes	
Ferro 2013	Middlebrook 7H10 (agar proportion)	Amikacin: 5.0	No	
		Kanamycin: 5.0	Yes	
FIND 2016	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 2.5	Yes	
		Capreomycin: 2.5	Yes	
Hillemann 2009	MGIT 960 and LJ (agar proportion)	Amikacin: 1.0 for MGIT and 40.0 for LJ	Yes for MGIT; no for LJ	
		Capreomycin: 2.5 for MGIT and 40.0 for LJ	Yes	Yes for both types of media
Huang 2011	Middlebrook 7H11 and MGIT 960	Amikacin: 1.0	Yes for MGIT; not applicable for 7H11	No WHO- recommended concentration specified for amikacin using 7H11
		Kanamycin: 6.0	Yes	
		Capreomycin: 10.0	Not applicable	
Ignatyeva 2012	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 5.0	No	
		Capreomycin: 2.5	Yes	
Jin 2013	LJ and BacT/ALERT 3D	Kanamycin: 10.0	Not applicable	No WHO- recommended concentrations specified for BacT/ALERT 3D media
		Capreomycin: unknown	Unknown	
Kiet	LJ	Kanamycin: 20.0	No	
Kontsevaya 2013	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 5.0	No	
		Capreomycin: 2.5	Yes	
Lacoma 2012	BACTEC 460TB	Kanamycin: 5.0	Not applicable	No WHO- recommended concentrations specified for BACTEC 460 media
		Capreomycin: 1.25	Not applicable	
Lopez-Roa 2012	Middlebrook 7H11 and MGIT 960	Amikacin: 4.0	No	

Miotto 2012	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 5.0	No	
		Capreomycin: 2.5	Yes	
NICD 2015	MGIT 960	Amikacin: 1.0	Yes	
		Kanamycin: 2.5	Yes	
		Capreomycin: 2.5	Yes	
Said 2012	Middlebrook 7H11	Kanamycin: 5.0	No	
		Capreomycin: 10.0	No	
Simons 2015	MGIT 960 and Middlebrook 7H10 (agar dilution)	Amikacin: 1.0 for MGIT and 5.0 for Middlebrook 7H10	Yes for MGIT; no for 7H10	
		Capreomycin: 2.5 for MGIT and 10.0 for Middlebrook 7H10	Yes for MGIT; no for 7H10	
Tagliani 2015	MGIT 960 and LJ (agar proportion)	Amikacin: 1.0 for MGIT and 30.0 for LJ	Yes for both types of media	
		Kanamycin: 2.5 for MGIT and 30.0 for LJ	Yes for both types of media	
		Capreomycin: 2.5 for MGIT and 40.0 for LJ	Yes for both types of media	
Tomasicchio 2016	MGIT 960	Amikacin: 1.0	Yes	
Tukvadze 2014	LJ	Kanamycin: 30.0	Yes	
		Capreomycin: 40.0	Yes	
van Ingen 2010	Middlebrook 7H10 (agar proportion)	Amikacin: 5.0	No	
		Capreomycin: 10.0	No	
Zivanovic 2012	MGIT 960 and LJ agar proportion	Amikacin: 1.0 for MGIT and 40.0 for LJ	Yes for MGIT; no for LJ	

		Capreomycin: 2.5 for MGIT and 40.0 for LJ	Yes for both types of media	

LJ-Löwenstein-Jensen; MGIT –Mycobacterial growth indicator tube

Reference: WHO, Updated interim critical concentrations for first-line and second-line DST (as of May 2012)
http://www.stoptb.org/wg/gli/assets/documents/Updated%20critical%20concentration%20table_1st%20and%202nd%20line%20drugs.pdf

Web Annex D.17. High complexity hybridization-based NAATs: Diagnostic accuracy for detection of resistance to pyrazinamide. A systematic review

Dr. Michael G Whitfield[1,2] and Prof. Dr. Annelies van Rie[1]

[1]Family Medicine and Population Health (FAMPOP), Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium; [2]Institute of Health Informatics Research, University College London, London, United Kingdom.

Corresponding author:
Michael G Whitfield, PhD
Institute of Health Informatics Research
222 Euston Road
University College London
London
NW1 2DA
United Kingdom
mwhit3007@gmail.com

This report has been prepared for the WHO Global TB Programme. Please do not distribute further.

BACKGROUND

Mycobacterium tuberculosis (MTB) remains in the top 10 of causes for death globally, with an estimated 10 million cases and 1.2 million deaths globally in 2019.[1] Of the 10 million cases, an estimated 465,000 were incident cases of multidrug resistant (MDR)/ rifampicin resistant (RR) tuberculosis, a marker for MDR-tuberculosis (TB).[1]

There have been great improvements in the detection of MTB and resistance to rifampicin (RIF) since the recommendation by WHO of the Xpert MTB/RIF and Xpert ULTRA MTB/RIF assays (Cepheid, Sunnyvale, USA) for routine diagnosis of MTB.[1-5] The other diagnostics recommended for routine use are the Hain GenoType MTBDR*plus* and GenoType MTBDR*sl* assays (Hain LifeScience, Nehren, Germany), which assist in the diagnosis of resistance to RIF and isoniazid or resistance to injectable drugs and fluoroquinolones, respectively.[1, 6] These two assays are based on hybridization-based technology, whereby a targeted gene (known to be associated with resistance to a specific antibiotic) is amplified, and probe technology is used to detect the presence of mutations which are known to be associated with resistance.[7]

Pyrazinamide (PZA) remains an important antibiotic for the treatment of both drug susceptible and drug resistant TB due to its unique ability to eradicate persister bacilli and its synergistic properties with other antibiotics.[8, 9] While mono-resistance to PZA is rare, PZA resistance is strongly associated with MDR/RR-TB, with an estimated 30-60% of MDR/RR-TB also resistant to PZA.[8-10] For people diagnosed with RR-TB, it is thus important to detect the presence of PZA resistance so that clinicians can make an informed decision on whether to include or exclude PZA in the treatment regimen.

Culture-based phenotypic drug susceptibility testing on the BACTED MGIT 960 system (Becton Dickinson Diagnostic Systems) is the current reference standard method to detect PZA resistance, but this method suffers from important technical challenges such as the needed for a specific pH level required and innolucum size.[11, 12] Consequently, PZA resistance is currently not assessed in routine practice. Genomic assays for determining of PZA requires assessment of the *pncA* gene (561bp), which is known to confer resistance to PZA.[13, 14] Currently, targeted or whole genome sequencing is used in research settings to identify mutations in the *pncA* gene. While there is a good association between genotype and phenotype results for PZA resistance (85-99% agreement), neither test is easily implemented into a routine setting.[8, 9, 15]

In 2007, Nipro (Osaka, Japan) developed GenoScholar PZA-TB, a line probe assay (LPA) with hybridization-based technology for detection of PZA resistance.[16] This assay is the first commercially available rapid molecular test for detection of PZA resistance. Compared to MTBDR*plus* and MTBDR*sl* LPA, the GenoScholar PZA-TB LPA does not include specific mutant probes, as resistance mutations are widespread across the entire *pncA* gene with no predominant mutations. Instead, the

GenoScholar PZA-TB assay targets a 700bp fragment that covers the entire *pncA* gene and promoter region up to nucleotide -18 of the wild type H37Rv reference strain.

DNA extracted from clinical specimens or cultures is amplified with primers by the polymerase chain reaction (PCR). Amplified DNA than hybridize to complementary oligonucleotides probes that are bound on a membrane-based strip. After allowing for hybridization of complimentary amplicons, alkaline phosphatase-labelled streptavidin is added to bind to any hybrids formed in the previous step. The enzymatic reaction with the substrate results in purple bands which are visually interpreted. The absence of wild type probe binding indicates the presence of a mutation. The first version of the assay contained 47 probes which cover the *pncA* promoter and open reading frame. The second version consisted of 48 probes. Three of the 48 probes (*pncA* 16, 17 and 35) in the second version represent silent mutations known to be a phylogenetic markers not associated with PZA resistance: Gly60Gly (probe 16), Ser65Ser (probe 17), and Thr142Thr (probe 35).

The molecular GenoScholar PZA-TB LPA assay, which is already commercially available, presents a method which could potentially be implemented for diagnosis of PZA resistance in routine care. However, limited data has been published on the diagnostic accuracy of the assay. This systematic review with meta-analysis aims to assist to collate all the available data to understand the diagnostic accuracy of the PZA LPA assay for detection of PZA resistance in TB patients in order to guide policy makers and clinicians.

Review objectives

The goal of the systematic literature review and meta-analysis is to estimate the diagnostic accuracy of the PZA LPA assay for detection of PZA resistance in cultured MTB isolates and sputum samples from patients diagnosed with pulmonary TB with or without rifampicin resistance using three different reference standards: phenotypic DST (pDST), genotypic DST (gDST) and a composite reference standard (CRS). The goal is to, if the data allows, to estimate the diagnostic accuracy of the PZA LPA overall and stratified by sputum smear microscopy status, rifampicin resistance status and treatment outcome.

METHODS

Eligibility criteria

Cross-sectional studies and cohort studies that assessed the diagnostic accuracy of the index test (GenoScholar PZA-TB LPA version 1 or 2) were eligible for inclusion. Both studies where the reference standard was performed after the index test and those where the reference standard was performed before the index test were eligible for inclusion. To be eligible for inclusion in the analysis, articles had to present data on PZA resistance detected by both the hybridization-based technology (PZA LPA) and by phenotypic culture-based DST and/or genotypic sequencing of the *pncA* gene. Studies were eligible

for inclusion is they reported data comparing the index test to an acceptable reference standard (defined below) from which we could extract true positive (TP), true negative (TN), false negative (FN) and false positive (FP) values.

We only included data on *Mycobacterium tuberculosis* and excluded any data on nontuberculous mycobacteria. We included results obtained from sputum samples or MTB cultures retrieved from people diagnosed pulmonary tuberculosis, independent of age, and HIV status. Samples were included irrespective of knowledge of rifampicin resistance (rifampicin resistance, present, absent or unknown). Studies performed at any types of health facilities and any level of laboratory levels from any country were eligible for inclusion.

Index test

Nipro GenoScholar PZA-TB LPA is the main index test in this review.

Target conditions

Tuberculosis

Reference standards

The microbiological reference standard is pDST performed using BD MGIT 960 PZA liquid assay, or another acceptable phenotypic based assay.

The microbiological reference standard is gDST performed using either targeted sequencing of the *pncA* gene or whole genome sequencing. We defined all samples with a *pncA* wild type to be susceptible, while any variant in *pncA* was considered resistant.

We defined the composite reference standard by classifying all samples with *pncA* wild type, *pncA* silent mutations and neutral mutations (Koser *et al*) to be susceptible, while any other variant in *pncA* was considered resistant.[17]

Outcomes

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Search methods

We searched the following databases: PubMed; Web of Science and EMBASE without language or date restrictions. We used the following search query: (PZA OR pyrazinamide OR pncA) AND (tuberculosis) AND ("line-probe assay" OR LPA OR "hybridization-based technology"). In addition, we approached Nipro (Osaka, Japan) to identify non-published data.

Data collection and analysis

Selection of studies

Two review authors independently assessed studies for eligibility. We resolved disagreements by discussion between the two reviewer authors. We illustrate the study selection process in a PRISMA diagram.

Data extraction

Two review authors independently extracted data from the included studies using a standardized form. Data was extracted using a standardized form and included first author name; publication year, publication title, PZA LPA result, phenotypic PZA DST result, genotypic *pncA* result, smear status, resistance profile, treatment information and any estimates of accuracy (sensitivity, specificity, positive and negative predictive value). We resolved disagreements by discussion with the two reviewer authors.

Assessment of methodological quality

Two review authors working independently assessed methodological quality using QUADAS-2 tailored to this review. We resolved disagreements by discussion between the two review authors.

Statistical analysis and data synthesis

We plotted estimates of the studies observed sensitivities and specificities in forest plots with 95% confidence intervals (CIs) using Review Manager (RevMan). Where adequate data were available (4 or more studies), we used pooling by summing up the numbers of TP, FP, TN and FN to calculate a pooled estimate of sensitivity and specificity. Where inadequate data (less than 4 studies) was available we simply show the sensitivity and specificity of available data, no pooling was conducted.

Assessment of certainty of the evidence (GRADE)

We assessed the certainty of the evidence using the GRADE approach, and the GRADEpro Guideline Development Tool (GDT) software (GRADEpro GDT 2015).[18, 19] For each outcome, we considered the certainty of the evidence as high when high-quality observational studies (cross-sectional or cohort studies) enrolled participants with diagnostic uncertainty. We used our judgment to downgrade the quality by one level if the reason was serious and two levels if the reason was very serious.[20, 21]

To apply GRADE, we used QUADAS-2 tool to assess risk of bias and applicability and made judgements with regarding to indirectness, inconsistency and impression.

Indirectness: we used QUADAS-2 tool for concerns of applicability and looked for important differences between the populations studied, the setting, index test, and the outcomes, and asked whether differences were sufficient to lower certainty in the results.

Inconsistency: we downgraded for unexplained inconsistency in sensitivity and specificity estimates.

Imprecision: we considered a precise estimate to be one that would allow a clinically meaningful decision. In addition, we calculated projected ranged for true positives (TP), false negatives (FN), true negatives (TN) and false positives (FP) for a given prevalence of pyrazinamide resistance and made

judgements on imprecision from these calculations. Prevalence's were selected based on published prevalence's of PZA resistance, which are mainly determined by the MTB drug resistance profile: 8% prevalence PZA-R for any person with TB; 50% for MDR/RR-TB and 90% for extensively drug resistant (XDR)-TB.[8-10, 22-24]

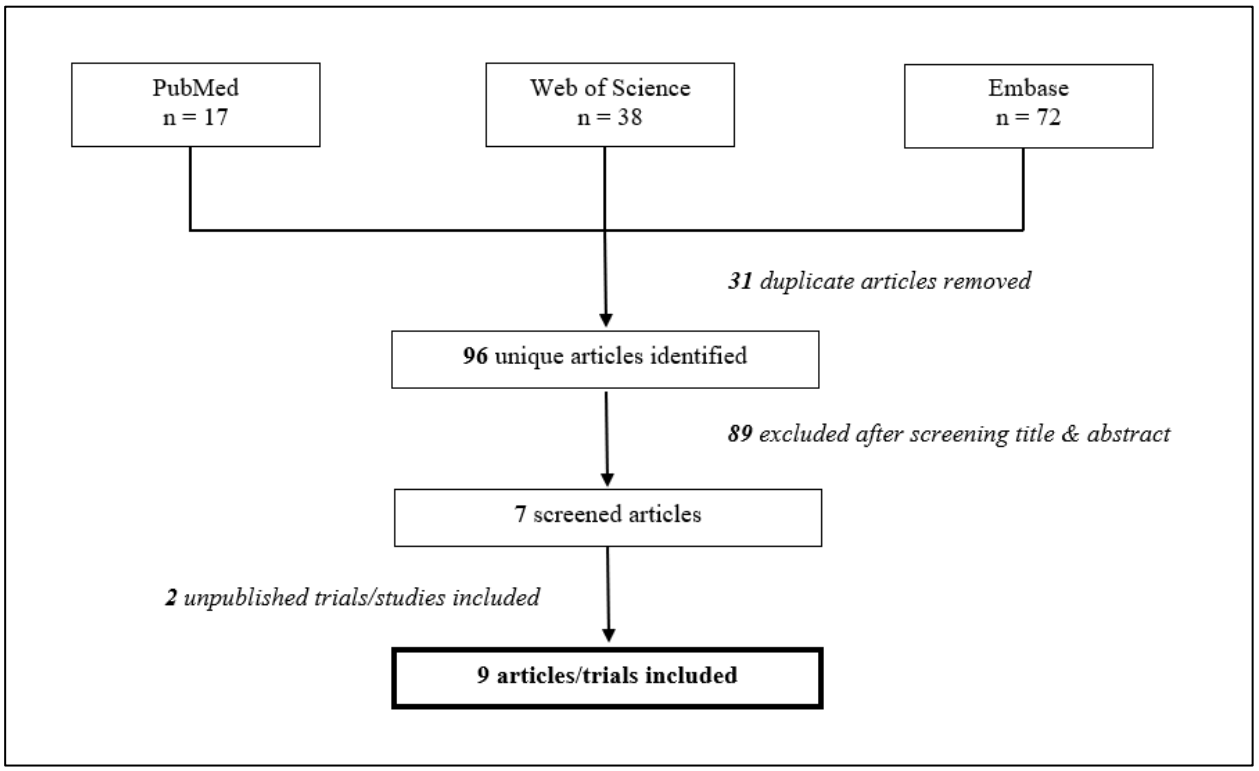
Publication bias: we rated publication bias as undetected (not serious) because of the comprehensiveness of the literature search and following extensive outreach to tuberculosis researchers to identify studies.

RESULTS

Identification of eligible studies

The initial search resulted in 96 unique records, from which 89 were not eligible and excluded from further analysis. After full-text review, the remaining 7 studies were retained for inclusion in the quantitative meta-analysis.[16, 25-30] In addition, we approached the principal investigator of 2 on-going trials evaluating the diagnostic accuracy of the Nipro PZA LPA assay: one at the Institute of Tropical Medicine in Belgium and the Persahabatan Hospital in Indonesia. The PRISMA diagram and reasons for exclusion are included below.[31]

Figure 1: PRISMA flow diagram and reasons for exclusion.



Exclusion reasons	Number of studies
No PZA LPA data	54

Review article	20
General article about TB (DR-TB)	9
Article studying NTMs	3
Commentary or editorials	3
Total	**89**

Quality assessment of studies

The results of the assessment of methodical quality using the QUADAS-2 tool is shown in Table 1 and Figure 2. For the patient selection domain, we judge 44% of studies to have high risk of bias because they selected specific isolates based on PZA resistance instead of a sample representative for the target population. For the index domain, we judged 44% studies to have a high risk of bias as they conducted the index test in an unblinded manner which increases the risk of bias when interpreting the index test results. For the reference standard domain, we judged most studies (89%) to have a low risk of bias as all studies except one used the standard reference methods (MGIT) to detect phenotypic PZA resistance. For the flow and timing domain, we judged that most (78%) studies were at low risk of bias as they conducted all tests on the same sample using the same reference standard and no patients/isolates were excluded.

Regarding applicability, in the patient selection domain, we judged 56% of studies to be of high concern due to patient selection not matching the review question. Instead, many studies aimed to test a large number of different *pncA* variants to the assay, thus not using a representative sample population. With respect to applicability of the index test, we judged all studies to be low concern based on the fact that we are interested in a one type of assay produced by one company. Finally, regarding the applicability of the reference standard, we consider most studies (89%) to be of low concern given the correct use of the standard reference methods used.

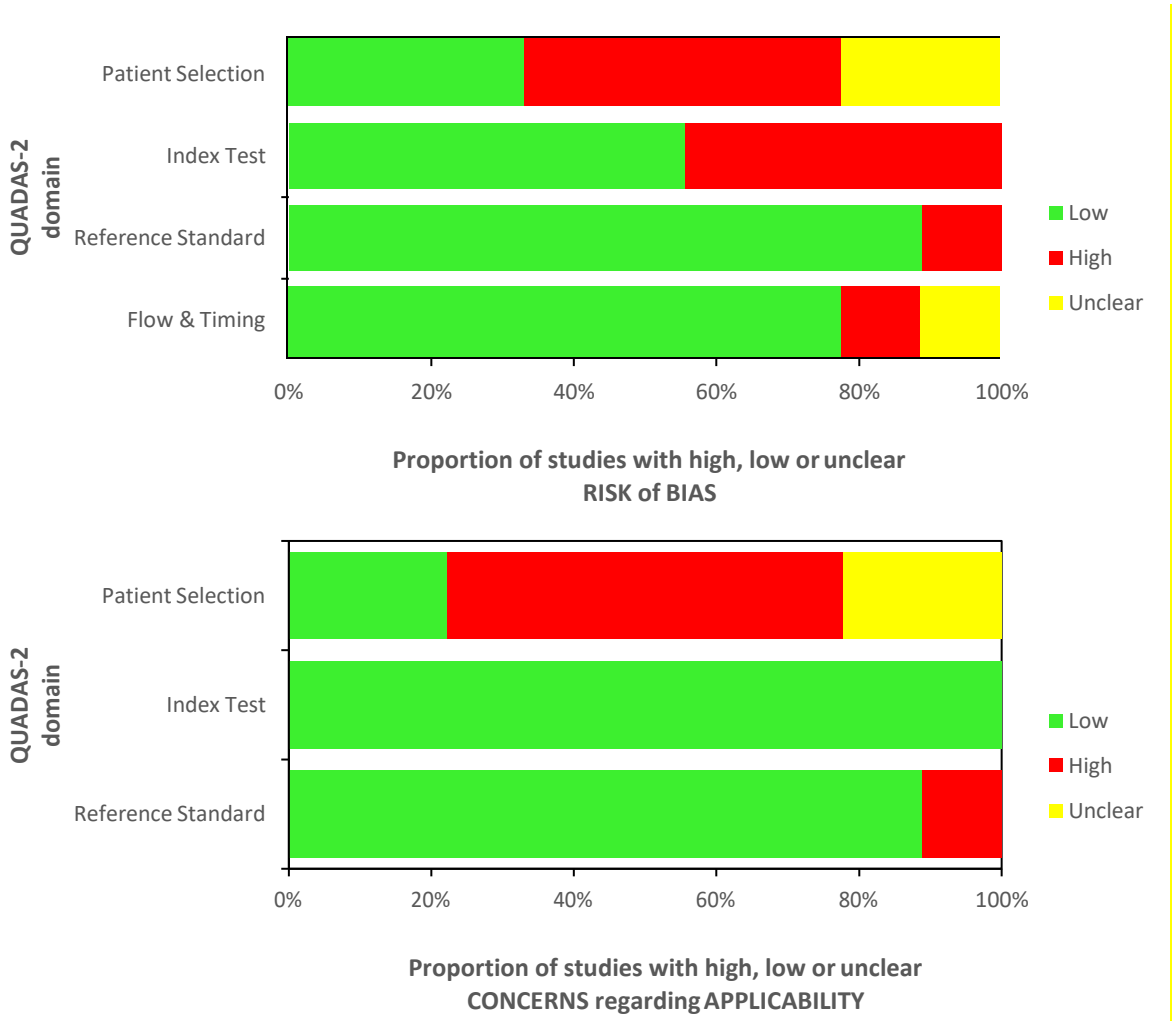
Table 1: Tabular presentation for QUADAS-2 results.

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference
Ando	-	-	+	+	-	+	+
Sekiguchi	?	-	+	+	?	+	+
Mitarai	+	+	-	?	?	+	-
Rienthong	?	-	+	-	-	+	+
Driesen	-	+	+	+	-	+	+
Willby	-	+	+	+	-	+	+
Matsumoto	+	-	+	+	+	+	+
ITM trial	-	+	+	+	-	+	+
Burhan trial	+	+	+	+	+	+	+

Risk/concern:

+ *Low* - *High* ? *Unclear*

Figure 2: Overview of QUADAS-2 assessment results.



PICO questions

PICO 1: What is the overall diagnostic accuracy of PZA LPA assay for detection of PZA resistance in sputum from PTM patients with and without RR-TB. Note: if sufficient data available: perform stratified analysis by DR or DS and by smear positive or smear negative.

PICO 2: What is the overall diagnostic accuracy of PZA LPA assay for detection of PZA resistance in isolates from PTM patients with and without RR-TB. Note: if sufficient data available: perform stratified analysis by DR or DS and by smear positive or smear negative.

GRADE questions

42) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, pDST?

No pooled estimates – too few studies to inform on certainty.

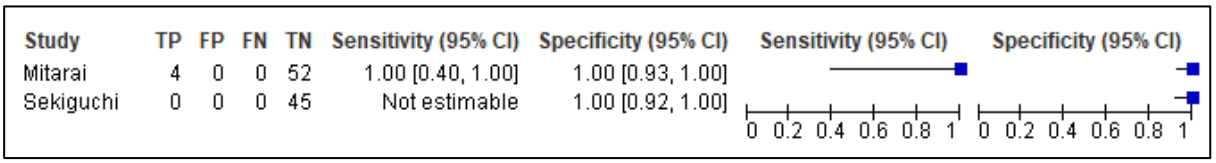


Figure 3: Forest plots of Nipro PZA LPA sensitivity and specificity for detection of PZA resistance against a reference standard of phenotypic drug susceptibility PZA assay (microbiological reference standard). The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

43) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, gDST?

No studies available to inform on decision.

44) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, CRS?

No studies available to inform on decision.

45) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, pDST?

No studies available to inform on decision.

46) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, gDST?

Only 1 study conducted where resistance profile of isolates is available – not enough evidence to make decision.

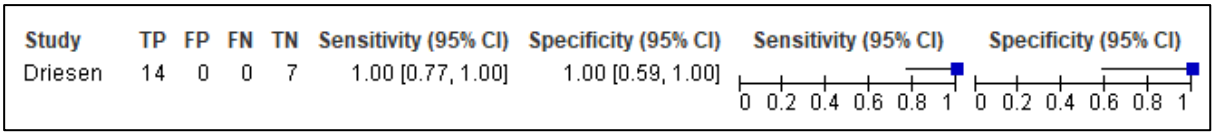


Figure 4: Forest plots of Nipro PZA LPA sensitivity and specificity for detection of PZA resistance against a reference standard of genomic drug susceptibility PZA assay (microbiological reference standard). The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

47) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, CRS?

No studies available to inform on decision.

Table 2: PICO 1 summary: diagnostic accuracy of PZA LPA for detection of PZA resistance in sputum from pulmonary TB patients with and without RR-TB.

PICO sub question	Reference Standard	Studies (participants)	Pooled Sensitivity (95% CI)	Pooled Specificity (95% CI)	Positive Predictive Value % (95% CI)	Negative Predictive Value % (95% CI)
Irrespective of rifampicin resistance						
42	pDST	2 (101)	Not estimable	Not estimable	Not estimable	Not estimable
43	gDST	0 (0)	NA	NA	NA	NA
44	CRS	0 (0)	NA	NA	NA	NA
With detected rifampicin resistance						
45	pDST	0 (0)	NA	NA	NA	NA
46	gDST	1 (21)	Not estimable	Not estimable	Not estimable	Not estimable
47	CRS	0 (0)	NA	NA	NA	NA

Abbreviations: CI: confidence interval; pDST: phenotypic drug susceptibility test; gDST: genomic drug susceptibility test; CRS: composite reference standard; NA: not applicable.

Table 3: GRADE certainty of evidence for sub PICO questions.

PICO sub question	Population	Reference Standard	Studies (participants)	Pooled Sensitivity (95% CI)	Pooled Specificity (95% CI)	Certainty of Evidence	Explanations
42	Irrespective of rifampicin resistance	pDST	2 (101)	Not estimable	Not estimable	NA	NA
43	Irrespective of rifampicin resistance	gDST	0 (0)	NA	NA	NA	NA
44	Irrespective of rifampicin resistance	CRS	0 (0)	NA	NA	NA	NA
45	With detected rifampicin resistance	pDST	0 (0)	NA	NA	NA	NA
46	With detected rifampicin resistance	gDST	1 (21)	Not estimable	Not estimable	NA	NA
47	With detected rifampicin resistance	CRS	0 (0)	NA	NA	NA	NA

Abbreviations: CI: confidence interval; pDST: phenotypic drug susceptibility test; gDST: genomic drug susceptibility test; CRS: composite reference standard; NA: not applicable.

48) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, pDST?

PZA LPA pooled sensitivity (95% CI) was 81.2% (75.4 to 85.8) and specificity (95% CI) was 97.8% (96.5 to 98.6), (7 studies, 964 participants; very low certainty evidence for sensitivity and low certainty evidence for specificity).

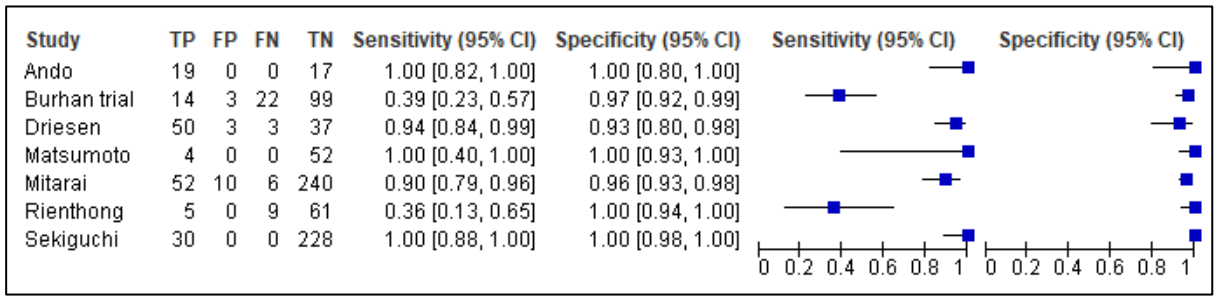


Figure 5: Forest plots of Nipro PZA LPA sensitivity and specificity for detection of PZA resistance against a reference standard of phenotypic drug susceptibility PZA assay (microbiological reference standard). The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

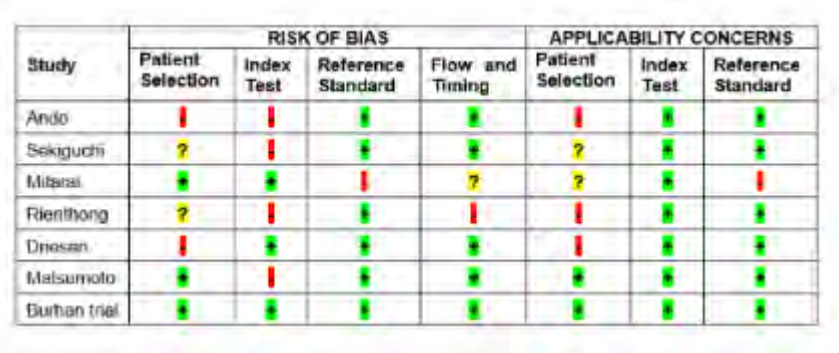


Figure 6. Risk of bias and applicability concerns for pDST as reference standard.

49) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, gDST?

PZA LPA pooled sensitivity (95% CI) and specificity (95% CI) were 96.4% (93.3 to 98.4) and 100% (99.4 to 100.0), (6 studies, 858 participants; low certainty evidence for sensitivity and low certainty evidence for specificity).

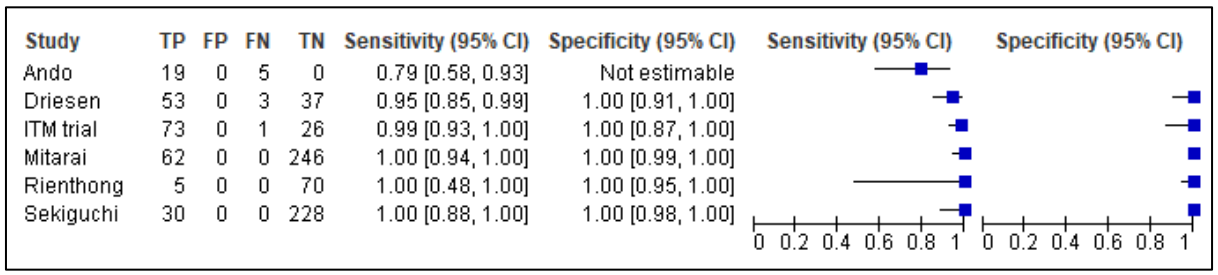


Figure 7: Forest plots of Nipro PZA LPA sensitivity and specificity for detection of PZA resistance against a reference standard of genomic drug susceptibility PZA assay (microbiological reference standard). The squares represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.

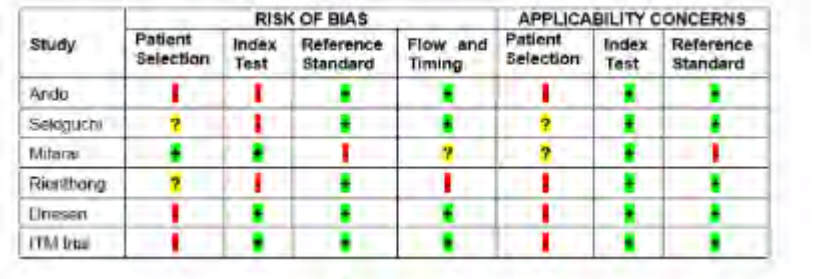


Figure 8. Risk of bias and applicability concerns for gDST as reference standard.

50) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, CRS?

PZA LPA pooled sensitivity (95% CI) and specificity (95% CI) were 95.7% (93.5 to 97.2) and 98.6% (97.4 to 99.2), (7 studies, 1140 participants; low certainty evidence for sensitivity and low certainty evidence for specificity).

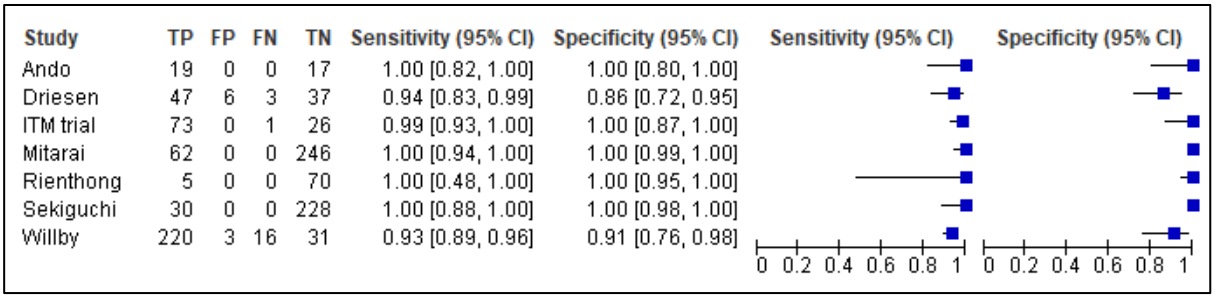
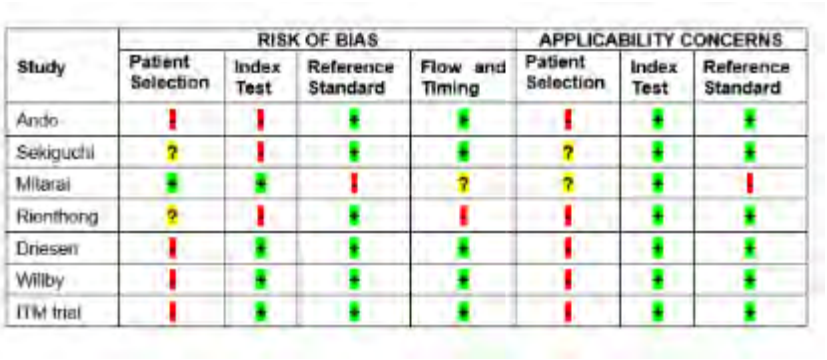


Figure 9: Forest plots of Nipro PZA LPA sensitivity and specificity for detection of PZA resistance against a composite reference standard of PZA resistance (composite reference standard). The squares

represent the sensitivity and specificity of one study, the black line its confidence interval. TP: true-positive; FP: false-positive; FN: false-negative; TN: true-negative.



Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Ando	!	!	+	+	!	+	+
Sekiguchi	?	!	+	+	?	+	+
Mitarai	+	+	!	?	?	+	!
Rienthong	?	!	+	!	!	+	+
Driesen	!	+	+	+	!	+	+
Willby	!	+	+	+	!	+	+
ITM trial	!	+	+	+	!	+	+

Figure 10. Risk of bias and applicability concerns for CRS as reference standard.

51) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, pDST?

No studies available to inform on decision.

52) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, gDST?

No studies available to inform on decision.

53) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, CRS?

No studies available to inform on decision.

Table 4: PICO 2 summary: diagnostic accuracy of PZA LPA for detection of PZA resistance in isolates from pulmonary TB patients with and without RIF resistant TB.

PICO sub question	Reference Standard	Studies (participants)	Sensitivity (95% CI)	Specificity (95% CI)	Posit Valu CI)
Irrespective of rifampicin resistance					
48	pDST	7 (964)	81.2 (75.4 to 85.8)	97.8 (96.5 to 98.6)	91.6
49	gDST	6 (858)	96.4 (93.3 to 98.4)	100 (99.4 to 100.0)	100 (
50	CRS	7 (1140)	95.7 (93.5 to 97.2)	98.6 (97.4 to 99.2)	98.1
With detected rifampicin resistance					
51	pDST	0 (0)	NA	NA	NA
52	gDST	0 (0)	NA	NA	NA
53	CRS	0 (0)	NA	NA	NA

Abbreviations: CI: confidence interval; pDST: phenotypic drug susceptibility test; gDST: genomic drug susceptibility test; CRS: composite reference standard; NA: not applicable.

Table 5: GRADE certainty of evidence for sub PICO questions.

PICO sub question	Population	Reference Standard	Studies (participants)	Pooled Sensitiv ity (95% CI)	Pooled Specific ity (95% CI)	Certainty of Evidence	Explanations
42	Irrespective of rifampicin resistance	pDST	7 (964)	81.2 (75.4 to 85.8)	97.8 (96.5 to 98.6)	Very low (sensitivity) Low (specificity)	Downgraded one level for risk of bias; one level for inconsistency and one level for imprecision (sensitivity) Downgraded one level for risk of bias and one level for inconsistency (specificity)
43	Irrespective of rifampicin resistance	gDST	6 (858)	96.4 (93.3 to 98.4)	100 (99.4 to 100.0)	Low (sensitivity) Low (specificity)	Downgraded one level for risk of bias and one level for inconsistency (sensitivity) Downgraded one level for risk of bias and one level for inconsistency(specificity)
44	Irrespective of rifampicin resistance	CRS	7 (1140)	95.7 (93.5 to 97.2)	98.6 (97.4 to 99.2)	Low (sensitivity) Low (specificity)	Downgraded one level for risk of bias and one level for inconsistency (sensitivity)

							Downgraded one level for risk of bias and one level for inconsistency(specificity)
45	With detected rifampicin resistance	pDST	0 (0)	NA	NA	NA	NA
46	With detected rifampicin resistance	gDST	0 (0)	NA	NA	NA	NA
47	With detected rifampicin resistance	CRS	0 (0)	NA	NA	NA	NA

Abbreviations: CI: confidence interval; pDST: phenotypic drug susceptibility test; gDST: genomic drug susceptibility test; CRS: composite reference standard; NA: not applicable.

SUMMARY OF MAIN RESULTS

1) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, pDST?

Too few studies to address this question.

2) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, gDST?

Too few studies to address this question.

3) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, CRS?

Too few studies to address this question.

4) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, pDST?

Too few studies to address this question.

5) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, gDST?

Too few studies to address this question.

6) Should PZA LPA assay on sputum be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, CRS?

Too few studies to address this question.

7) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, pDST?

Pooled sensitivity was 81.2% (75.4 to 85.8) and specificity was 97.8% (96.5 to 98.6)
Certainty of evidence was very low for sensitivity and low specificity.

8) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, gDST?

Pooled sensitivity was 96.4% (93.3 to 98.4) and specificity was 100% (99.4 to 100)
Certainty of evidence was low for both sensitivity and specificity.

9) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, irrespective of resistance to RIF, CRS

Pooled sensitivity was 95.7% (93.5 to 987.2) and specificity was 98.6% (97.4 to 99.2)
Certainty of evidence was low for both sensitivity and specificity.

10) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, pDST?

Too few studies to address this question.

11) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, gDST?

Too few studies to address this question.

12) Should PZA LPA assay on isolates be used to diagnose PZA resistance in patients with microbiologically confirmed PTB, with detected resistance to RIF, CRS?

Too few studies to address this question.

AUTHORS CONCLUSIONS

The Nipro Genoscholar PZA TB assay appears to perform accurately to detect the presence of a wide range of different *pncA* variants. Majority of the studies conducted the assay using stored culture isolates; thus, we are unable to reach any conclusions regarding the use of the assay in clinical specimens. None of the studies stratified the results by resistance status, smear status or treatment outcome and so we are unable to reach any conclusions in this regard.

IMPLICATIONS FOR RESEARCH

This review suggests that the Nipro Genoscholar PZA TB assay is accurate with the detection of variants in the *pncA* gene. Due to the large number of variants identified and subsequently be showed to be associated with phenotypic PZA resistance, major of studies selected a wide range of different *pncA* variants to run on the assay. Thus, it appears that the assay is a good diagnostic which could potentially be implemented in routine care/laboratory settings. We would also recommend studies investigate the performance of the assay in clinical specimens (sputum) and not from stored culture isolates to truly determine the clinical usefulness from routine settings.

However, we would recommend that there are cross-sectional and/or cohort studies conducted which investigates the true diagnostic accuracy of the assay in routine care. These studies would provide strong quality of evidence which could be used to better improve on the use of the assay in routine settings.

Given the use of PZA in DR treatment regimens, we would suggest that the PZA LPA be conducted as a reflex diagnostic upon the detection of rifampicin resistance from a routine initial diagnostic such as Xpert MTB/RIF or Ultra MTB/RIF.

This report has been prepared for the WHO Global TB Programme. Please do not distribute further.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the international TORCH consortium for their valuable comments.

We thank Erlina Burhan from Persahabatan Hospital in Indonesia for sharing data from the ongoing trial.

We thank Leen Rigouts and Bouke de Jong from Institute of Tropical Medicine (ITM) in Antwerp, Belgium for sharing data from their finalized trial.

CONTRIBUTIONS OF AUTHORS

Both MGW and AVR were involved in all processes of this report.

DECLARATIONS OF INTEREST

MGW received funding from the World Health Organization Global Tuberculosis Programme, Geneva.

AVR received funding from the World Health Organization Global Tuberculosis Programme, Geneva.

REFERENCES

1. World Health Organization. Global Tuberculosis Report 2020. https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf?ua=1.
2. World Health Organization. Xpert MTB/RIF implementation manual Technical and operational 'how-to': practical considerations 2014. http://apps.who.int/iris/bitstream/handle/10665/112469/9789241506700_eng.pdf?sequence=1.
3. World Health Organization. Rapid Implementation of the Xpert MTB/RIF diagnostic test Technical and Operational 'How-to'Practical considerations 2011. http://apps.who.int/iris/bitstream/handle/10665/44593/9789241501569_eng.pdf?sequence=1.
4. World Health Organization. Monitoring of Xpert MTB/RIF roll-out 2016. https://www.who.int/tb/areas-of-work/laboratory/mtb-rif-rollout/en/.
5. Chakravorty S, Simmons AM, Rowneki M et al. The New Xpert MTB/RIF Ultra: Improving Detection of *Mycobacterium tuberculosis* and Resistance to Rifampin in an Assay Suitable for Point-of-Care Testing. *mBio* 2017; **8**: e00812-17.
6. World Health Organization. Molecular Line Probe Assays for Rapid Screening of Patients at Risk of Multidrug-Resistsant Tuberculosis (MDR-TB) - Policy Statement 2008. http://www.who.int/tb/features_archive/policy_statement.pdf?ua=1.
7. Hain Lifescience. Genotype MTBDR*plus* VER 2.0 Instructions for Use 2015. https://www.ghdonline.org/uploads/MTBDRplusV2_0212_304A-02-02.pdf.
8. Whitfield MG, Soeters HM, Warren RM et al. A Global Perspective on Pyrazinamide Resistance: Systematic Review and Meta-Analysis. *PLoS One* 2015; **10**: e0133869.
9. Chang KC, Yew WW, Zhang Y. Pyrazinamide Susceptibility Testing in *Mycobacterium tuberculosis*: a Systematic Review with Meta-Analyses. *Antimicrob Agents Chemother* 2011; **55**: 4499-505.
10. Whitfield MG, Streicher EM, Dolby T et al. Prevalence of pyrazinamide resistance across the spectrum of drug resistant phenotypes of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 2016; **99**: 128-30.
11. Chedore P, Bertucci L, Wolfe J et al. Potential for Erroneous Results Indicating Resistance When Using the Bactec MGIT 960 System for Testing Susceptibility of *Mycobacterium tuberculosis* to Pyrazinamide. *J Clin Microbiol* 2010; **48**: 300-1.
12. Whitfield MG, Warren RM, Streicher EM et al. *Mycobacterium tuberculosis pncA* Polymorphisms That Do Not Confer Pyrazinamide Resistance at a Breakpoint Concentration of 100 Micrograms per Milliliter in MGIT. *J Clin Microbiol* 2015; **53**: 3633-5.
13. Scorpio A, LindholmLevy P, Heifets L et al. Characterization of *pncA* mutations in pyrazinamide-resistant *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 1997; **41**: 540-3.
14. Scorpio A, Zhang Y. Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat Med* 1996; **2**: 662-7.
15. Streicher EM, Maharaj K, York T et al. Rapid Sequencing of the *Mycobacterium tuberculosis pncA* Gene for Detection of Pyrazinamide Susceptibility. *J Clin Microbiol* 2014; **52**: 4056-7.
16. Sekiguchi J, Nakamura T, Miyoshi-Akiyama T et al. Development and evaluation of a line probe assay for rapid identification of pncA mutations in pyrazinamide-resistant *mycobacterium tuberculosis* strains. *J Clin Microbiol* 2007; **45**: 2802-7.
17. Köser CU, Cirillo DM, Miotto P. How To Optimally Combine Genotypic and Phenotypic Drug Susceptibility Testing Methods for Pyrazinamide. *Antimicrob Agents Chemother* 2020; **64**: e01003-20.

18. Schünemann HJ, Mustafa R, Brozek J et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol* 2016; **76**: 89-98.
19. Schünemann HJ, Oxman AD, Brozek J et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; **336**: 1106-10.
20. Schünemann HJ, Mustafa RA, Brozek J et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. *J Clin Epidemiol* 2020; **122**: 129-41.
21. Schünemann HJ, Mustafa RA, Brozek J et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. *J Clin Epidemiol* 2020; **122**: 142-52.
22. Khan MT, Malik SI, Ali S et al. Prevalence of Pyrazinamide Resistance in Khyber Pakhtunkhwa, Pakistan. *Microbial drug resistance (Larchmont, NY)* 2018; **24**: 1417-21.
23. Calderón RI, Velásquez GE, Becerra MC et al. Prevalence of pyrazinamide resistance and Wayne assay performance analysis in a tuberculosis cohort in Lima, Peru. *Int J Tuberc Lung Dis* 2017; **21**: 894-901.
24. Pang Y, Zhang Z, Wang Y et al. Genotyping and Prevalence of Pyrazinamide- and Moxifloxacin-Resistant Tuberculosis in China, 2000 to 2010. *Antimicrob Agents Chemother* 2017; **61**.
25. Ando H, Mitarai S, Kondo Y et al. Pyrazinamide resistance in multidrug-resistant *Mycobacterium tuberculosis* isolates in Japan. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 2010; **16**: 1164-8.
26. Driesen M, Kondo Y, de Jong BC et al. Evaluation of a novel line probe assay to detect resistance to pyrazinamide, a key drug used for tuberculosis treatment. *Clin Microbiol Infect* 2018; **24**: 60-4.
27. Willby MJ, Wijkander M, Havumaki J et al. Detection of *Mycobacterium tuberculosis pncA* Mutations by the Nipro Genoscholar PZA-TB II Assay Compared to Conventional Sequencing. *Antimicrob Agents Chemother* 2017; **62**: e01871-17.
28. Matsumoto T, Ogata H, Toyota E et al. [Clinical evaluation of a line probe assay kit for the identification of Mycobacterium species and detection of drug-resistant *Mycobacterium tuberculosis*]. *Kekkaku* 2013; **88**: 291-6.
29. Mitarai S, Kato S, Ogata H et al. Comprehensive multicenter evaluation of a new line probe assay kit for identification of Mycobacterium species and detection of drug-resistant *Mycobacterium tuberculosis*. *J Clin Microbiol* 2012; **50**: 884-90.
30. Rienthong S, Boonin C, Chaiyasirinrote B et al. Evaluation of a novel line-probe assay for genotyping-based diagnosis of *Mycobacterium tuberculosis* in Thailand. *Int J Tuberc Lung Dis* 2015; **19**: 817-22.
31. Moher D, Liberati A, Tetzlaff J et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 2009; **6**.

APPENDICES

Appendix 1. QUADAS-2 tool

Domain 1: Patient Selection

Risk of Bias: Could the selection of patients/specimens have introduced bias?

Signalling question 1: Was a consecutive or random sample of patients or specimens enrolled?

'yes' when the study enrolled a consecutive or random sample;

'no' for all other studies;

'unclear' when the study did not report on patient selection.

Signalling question 2: Was a case-control design avoided?

'yes' when a prospective or cross-sectional design was used;

'no' when a case-control design was used;

'unclear' for all other study designs.

Signalling question 3: Did the study avoid inappropriate exclusions?

'yes' when no indications of inappropriate exclusions were noted;

'no' if the study excluded samples based on characteristics such as a prior testing;

'unclear' not applicable.

Applicability: Are there concerns that the included patients and setting do not match the review question?

We were interested in compiling all available information on the diagnostic accuracy for hybridization-based technology which is commercially available for detection of pyrazinamide resistance. Since our goal was to compile as much data as possible, we do not expect any concerns of applicability concerning patient selection. Therefore, we judged applicability to be 'unclear' if no information on sampling was reported and judged applicability to be 'low concern' for all studies which reported on their sampling method.

Domain 2: Index Test

Risk of Bias: Could the conduct or interpretation of the index test have introduced bias?

Signalling question 1: Were the index test results interpreted with knowledge of the results of the reference standard?

'yes' when PZA LPA were done in a blinded manner;

'no' when PZA LPA were not done in a blinded manner;

'unclear' if not reported.

Signalling question 2: If a threshold was used, was it prespecified?

There is no threshold for the PZA LPA assay, the result is binary: visible band or not. We answered 'yes' for all studies.

Applicability: Are there concerns that the index test, its conduct, or its interpretation differ from the review question?

There are a limited number of articles available for this review question, and the index test comes from a single manufacturer. If the study used multiple or blinded readers for the interpretation of the PZA LPA we deemed the applicability as 'low concern'. Where a study used a single, unblinded reader for the interpretation of the PZA LPA we deemed the applicability as 'high concern' and for all other studies we deemed applicability as 'unclear'.

Domain 3: Reference Standard

Risk of Bias: Could the reference standard, its conduct, or its interpretation have introduced bias?

Signalling question 1: Is the reference standard likely to correctly classify the target condition?

'yes' when the study used an WHO-recommended or internationally recognized methods;

'no' for none standard methods;

'unclear' when not enough information was provided regarding the methods.

Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?

Limited room for interpretation on these assays – "yes" for all in this instance

'yes' when the index test was performed after the reference standard;

'no' when the index test was performed before the reference standard;

'unclear' when not reported.

Applicability: Are there concerns that the target condition as defined by the reference standard does not match the question?

There are several methods available for genotyping a clinical Mycobacterium tuberculosis isolate, such as whole genome sequencing, targeted (Sanger) sequencing and next generation sequencing. There are also several protocols and software packages which are available for the interpretation of genotypic data. In terms of phenotypic methods, most studies used the accepted/recognized standard method (MGIT 960) however, there is concerns regarding the reliability of this accepted standard as well due to the technical challenges. When this was not reported, we answered 'unclear'; when non-standard methods were used, we answered "high concern' and finally for all other studies, we answered 'low concern'.

Domain 4: Flow and Timing

Risk of Bias: Could the patient flow have introduced bias?

Signalling question 1: Was there an appropriate interval between the index test and reference standard?

This is not relevant for review question; 'yes' for all studies.

Signalling question 2: Did all patients receive the same reference standard?

‘yes’ when the same reference standard was used;
‘no’ when the samples were processed in different laboratories;
‘unclear’ not applicable.
Signaling question 3: Were all patients included in the analysis?
‘yes’ when all patients/isolates were included in the analysis;
‘no’ when not all patients/isolates were included in the analysis;
‘unclear’ not applicable.

This report has been prepared for the WHO Global TB Programme.

Web Annex D.18. Review of the diagnostic accuracy of targeted next-generation sequencing technologies for detection of drug resistance among people diagnosed with TB

Phu Phan Trieu[1], Eli Harriss[2], Timothy M Walker[1,3]

[1] Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
[2] Bodleian Health Care Libraries, University of Oxford, Oxford, UK
[3] Nuffield Department of Medicine, University of Oxford, Oxford, UK

Correspondence:
Timothy M Walker
twalker@oucru.org
timothy.walker@ndm.ox.ac.uk

Acknowledgements

We would like to thanks the following for support and advice in the generation of this document: Cecily Miller, Nazir Ismail, Mathieu Bastard, Matteo Zignol, Mikashmi Kohli, Samuel Schumacher, Tamara Kredo and Karen Steingart.

We would also like to thank all the contributors of data, all of whom we corresponded with regularly to ensure data accuracy. In no particular order: Mohammad Khaja Mafij Uddin, Faridath Massou, Leen Rigouts, Conor Meehan, Corinne Merle, Stefan Niemann, Silke Feuerriegel, Isabelle Bonnet, Tim Rodwell, Rebecca Coleman, Marva Seifert, Anita Suresh, Priti Kambli, Camilla Rodriques, Shaheed Vally Omar, Eric Wang, and Andrea Cabibbe, along with the wider teams they represent and who generated these data.

Declarations of interest

PPT is employed through Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam, with salary support from the Bill and Melinda Gates Foundation. EH is employed through the University of Oxford. TMW has received research funds from the World Health Organization, FIND, the Bill and Melinda Gates Foundation, the MRC (UK), IMI (European Union) and Wellcome. TMW is a Wellcome Clinical Career Development Fellow (214560/Z/18/Z), employed through the University of Oxford.

Executive summary

Molecular diagnostic tests for drug susceptibility testing (DST) of *Mycobacterium tuberculosis* strains are now widely understood to be the best hope for attaining the WHO's goal of universal access to drug susceptibility testing. WHO recommended rapid diagnostic (WRD) tests detect resistance to only a few select drugs due to the limited number of mutations that can be probed. Some national TB programmes have instituted whole-genome sequencing as a routine molecular test, but this approach remains culture dependent. Targeted next-generation sequencing (tNGS) technologies can be applied directly from clinical samples, avoiding the need for culture, and have the scope to target nearly all known resistance-associated mutations for clinically relevant drugs.

This systematic review of the diagnostic accuracy of tNGS platforms explores the performance of these solutions both as a test of the initial detection of drug resistance from sputum samples, and as a reflex test applied to samples already identified as rifampicin resistant by molecular WRDs. This review is to inform a Guideline Development Meeting considering the use of tNGS technologies, convened by the World Health Organization (WHO). We sought to include data from all commercially available platforms we are aware of: Deeplex® Myc-TB (GenoScreen); Deepchek®-TB RpoB/InhA Drug Resistance Assay (118A24) (Advanced Biological Laboratories); NanoTB® (Oxford Nanopore Technologies); and TBseq® (ShengTing Biotech). We obtained published data from a systematic review of the peer-reviewed literature and unpublished data obtained passively through a WHO call for data, or actively by contacting experts it the field. We also searched the product websites for additional references.

Most data originated from the Unitaid sponsored Seq & Treat project granted to FIND, who conducted a trial of three diagnostic platforms, namely Deeplex® Myc-TB, Deepchek®-TB RpoB/InhA and NanoTB®, on the same clinical samples collected in Georgia, South Africa and India. Of these, data from the Deepchek®-TB RpoB/InhA platform were not available in time for this review. In addition to data from FIND, data were contributed for Deeplex® Myc-TB from the ongoing Diagnostic of Multidrug-resistant tuberculosis in Africa (DIAMA) study being conducted across West and East Africa; from an ongoing study run by the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b); and from pilot projects run by the National Institute for Communicable Infections (NICD) in South Africa, and the San Raffaele Scientific Institute (SRSI) in Italy. Four papers were identified from the published literature, all assessing Deeplex® Myc-TB. These studies were based in Germany, France, and two in India. Data on the fourth platform, TBseq®, were generated independently by the Beijing Chest Hospital, but contributed by the assay manufacturer ShengTing Biotech in China.

An individual patient data (IPD) meta-analysis was performed to generate pooled sensitivity and specificity. Minimum thresholds of 100 susceptible and 50 resistant samples were set as inclusion criteria for each drug, although new and repurposed drugs (bedaquiline, clofazimine and linezolid) were not subject to these thresholds. For each drug, the best performing tNGS platform was included, along with any other platform with a midpoint estimate within 10% of the best platform's sensitivity and within 5% of its specificity. Phenotypic DST (pDST) was used as the reference standard for all drugs with exceptions for rifampicin, ethambutol and pyrazinamide for which separate analyses were conducted, one using pDST and one using a composite reference standard. For ethambutol and pyrazinamide this was a composite of pDST and whole-genome sequencing (WGS), whereas for rifampicin this was a composite of pDST, WGS and the rifampicin resistance results from either MTB/RIF Xpert® or Xpert Ultra®. All results were taken as given by the data contributors, or clarifications sought from the contributors where required. For Deeplex® Myc-TB, the presence of resistant alleles at any frequency was interpreted as a 'resistant' (R) result.

The IPD meta-analysis used a multi-variable, mixed-effects approach. Separate models were run for sensitivity and specificity, with these each coded as binary dependent variables. Additional co-variables were included as fixed effects: MTB/RIF Xpert® or Xpert Ultra® semi-quantitative result defined as either very low or low, or as medium or high; MTB/RIF Xpert® or Xpert Ultra® rifampicin

resistance result; whether a sample had been sequenced on one or on two tNGS platform (all the samples from FIND were sequenced using both Deeplex® Myc-TB and NanoTB®). Data on HIV test results were available only from FIND, where they were generated as part of the study, and were only included in models addressing the specific sub-analysis looking at diagnostic accuracy in people living with HIV infection. Study sites, including different countries within a single study, were included as a random effect in the models.

There were two PICO (population, intervention, comparator and outcome) questions. The first PICO (PICO1) asked if tNGS should be used as an initial test for drug resistance in patients with bacteriologically confirmed TB, for current first-line drugs (isoniazid, rifampicin, ethambutol, pyrazinamide), plus the fluoroquinolones levofloxacin and moxifloxacin. The second PICO (PICO2) asked if tNGS should be used in patients with bacteriologically confirmed rifampicin-resistant TB, for the drugs mentioned in PICO1, with the exception of rifampicin, plus two injectable drugs (streptomycin and amikacin), and the so-called new or repurposed drugs (bedaquiline, linezolid and clofazimine).

We used QUADAS-2 to assess any concerns of bias and applicability. We marked most studies as high risk for bias and applicability in the patient selection domain for PICO 1 due to the enrichment for rifampicin-resistant samples. Other concerns were that some studies used a non-WHO critical concentration for pDST to one or more drugs, and in two studies, the laboratory staff were not blinded to the index or reference test result when processing the other. In addition, one study used different samples for the index and reference tests. Where studies were enriched for rifampicin resistance, we downgraded our assessment of the quality of evidence for indirectness, but not for risk of bias. Where studies using a non-WHO recommended critical concentration, we downgraded our assessment of the quality of evidence for risk of bias, but not indirectness. We downgraded our assessments of the quality of evidence in all other instances where a relevant QUADAS-2 domain was assessed as high-risk.

The results of the IPD meta-analysis and grading of the evidence were as follows:

PICO 1

Rifampicin – composite reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 93% (87-99) and 96% (89-100) respectively, based on up to 9 studies and up to 1436 samples and an 84% prevalence of rifampicin resistance. The indeterminate rate was 12%. Quality of evidence was judged to be moderate for sensitivity and low for specificity.

Results indicate that, in theory, of 1000 patients, among whom 100 have rifampicin resistance, 93 would be detected (true positive) by tNGS and 7 would not be detected by tNGS (false negative); whereas 864 without rifampicin resistance would be negative by tNGS (true positive) and 36 would be reported as rifampicin resistant by tNGS (false positive).

Rifampicin – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 99% (97-100) and 81% (69-93) respectively, based on up to 13 studies and up to 961 samples and a 69% prevalence of rifampicin resistance. The indeterminate rate was 10%. Quality of evidence was judged to be moderate for sensitivity and very low for specificity.

Results indicate that, in theory, of 1000 patients among whom 100 have rifampicin resistance, 99 would be detected (true positive) by tNGS and 1 would not be detected by tNGS (false negative); whereas 729 without rifampicin resistance would be negative by tNGS (true positive) and 171 would be reported as rifampicin resistant by tNGS (false positive). However, as phenotypic DST for rifampicin is known to miss borderline-resistant strains, using the composite reference standard for this drug is preferred.

Isoniazid – pDST as reference standard.

Pooled sensitivity and specificity (95% confidence interval) were 96% (93-99) and 97% (95-99) respectively, based on up to 12 studies and up to 1440 samples and a 74% prevalence of isoniazid resistance. The indeterminate rate was 15%. Quality of evidence was judged to be moderate for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 100 have isoniazid resistance, 96 would be detected (true positive) by tNGS and 4 would not be detected by tNGS (false negative); whereas 873 without isoniazid resistance would be negative by tNGS (true positive) and 27 would be reported as isoniazid-resistant by tNGS (false positive).

Levofloxacin – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 94% (88-100) and 96% (93-99) respectively, based on up to 7 studies and up to 913 samples and a 42% prevalence of levofloxacin resistance. The indeterminate rate was 9%. Quality of evidence was judged to be low for both sensitivity and moderate for specificity.

Results indicate that, in theory, of 1000 patients among whom 50 have levofloxacin resistance, 47 would be detected (true positive) by tNGS and 3 would not be detected by tNGS (false negative); whereas 912 without levofloxacin resistance would be negative by tNGS (true positive) and 38 would be reported as levofloxacin resistant by tNGS (false positive).

Moxifloxacin – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 96% (92-99) and 96% (93-100) respectively, based on up to 8 studies and up to 921 samples and a 41% prevalence of moxifloxacin resistance. The indeterminate rate was 9%. Quality of evidence was judged to be moderate for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 50 have moxifloxacin resistance, 48 would be detected (true positive) by tNGS and 2 would not be detected by tNGS (false negative); whereas 912 without moxifloxacin resistance would be negative by tNGS (true positive) and 38 would be reported as moxifloxacin resistant by tNGS (false positive).

Pyrazinamide – composite as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 88% (85-92) and 99% (97-100) respectively, based on up to 3 studies and up to 364 samples and a 56% prevalence of pyrazinamide resistance. The indeterminate rate was 18%. Quality of evidence was judged to be moderate for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have pyrazinamide resistance, 26 would be detected (true positive) by tNGS and 4 would not be detected by tNGS (false negative); whereas 960 without pyrazinamide resistance would be negative by tNGS (true positive) and 10 would be reported as pyrazinamide resistant by tNGS (false positive).

Pyrazinamide – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 85% (80-90) and 94% (92-96) respectively, based on up to 6 studies and up to 425 samples and a 52% prevalence of pyrazinamide resistance. The indeterminate rate was 15%. Quality of evidence was judged to be low for both sensitivity and moderate for specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have pyrazinamide resistance, 26 would be detected (true positive) by tNGS and 4 would not be detected by tNGS (false negative); whereas 912 without pyrazinamide resistance would be negative by tNGS (true positive) and 58 would be reported as pyrazinamide resistant by tNGS (false positive). However, as phenotypic DST for pyrazinamide is known to be an imperfect reference standard, using the composite reference standard for this drug is preferred.

Ethambutol – composite as reference standard.

Pooled sensitivity and specificity (95% confidence interval) were 96% (94-99) and 99% (98-100) respectively, based on up to 4 studies and up to 432 samples and a 62% prevalence of ethambutol resistance. The indeterminate rate was 16%. Quality of evidence was judged to be low for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have ethambutol resistance, 29 would be detected (true positive) by tNGS and 1 would not be detected by tNGS (false negative); whereas 960 without ethambutol resistance would be negative by tNGS (true positive) and 10 would be reported as ethambutol resistant by tNGS (false positive).

Ethambutol – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 88% (82-94) and 94% (91-97) respectively, based on just 1 study and up to 334 samples and a 23% prevalence of ethambutol resistance. The indeterminate rate was 16%. Quality of evidence was judged to be low for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have ethambutol resistance, 26 would be detected (true positive) by tNGS and 4 would not be detected by tNGS (false negative); whereas 912 without ethambutol resistance would be negative by tNGS (true positive) and 58 would be reported as ethambutol resistant by tNGS (false positive). However, as phenotypic DST for ethambutol is known to miss borderline-resistant strains, using the composite reference standard for this drug is preferred.

PICO 2

Isoniazid – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 96% (94-99) and 96% (92-100) respectively, based on 12 studies and up to 1440 samples and a 74% prevalence of isoniazid resistance. The indeterminate rate was 15%. Quality of evidence was judged to be high for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 750 have isoniazid resistance, 720 would be detected (true positive) by tNGS and 30 would not be detected by tNGS (false negative); whereas 240 without isoniazid resistance would be negative by tNGS (true positive) and 10 would be reported as isoniazid resistant by tNGS (false positive).

Levofloxacin – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 96% (90-100) and 96% (93-99) respectively, based on up to 7 studies and up to 913 samples and a 42% prevalence of levofloxacin resistance. The indeterminate rate was 9%. Quality of evidence was judged to be moderate for sensitivity and high for specificity.

Results indicate that, in theory, of 1000 patients among whom 300 have levofloxacin resistance, 288 would be detected (true positive) by tNGS and 12 would not be detected by tNGS (false negative); whereas 672 without levofloxacin resistance would be negative by tNGS (true positive) and 28 would be reported as levofloxacin resistant by tNGS (false positive).

Moxifloxacin – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 97% (94-100) and 95% (91-99) respectively, based on up to 8 studies and up to 921 samples and a 41% prevalence of moxifloxacin resistance. The indeterminate rate was 9%. Quality of evidence was judged to be high for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 300 have moxifloxacin resistance, 291 would be detected (true positive) by tNGS and 9 would not be detected by tNGS (false negative); whereas 665 without moxifloxacin resistance would be negative by tNGS (true positive) and 35 would be reported as moxifloxacin resistant by tNGS (false positive).

Pyrazinamide – composite as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 90% (87-93) and 99% (97-100) respectively, based on 3 studies and up to 346 samples and a 56% prevalence of pyrazinamide resistance. The indeterminate rate was 18%. Quality of evidence was judged to be high for both sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 500 have pyrazinamide resistance, 450 would be detected (true positive) by tNGS and 50 would not be detected by tNGS (false negative); whereas 495 without pyrazinamide resistance would be negative by tNGS (true positive) and 5 would be reported as pyrazinamide resistant by tNGS (false positive).

Pyrazinamide – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 90% (85-95) and 90% (86-94) respectively, based on 6 studies and up to 425 samples and a 53% prevalence of pyrazinamide resistance. The indeterminate rate was 15%. Quality of evidence was judged to be moderate for sensitivity and high for specificity. However, as phenotypic DST for pyrazinamide is known to be an imperfect reference standard, using the composite reference standard for this drug is preferred.

Results indicate that, in theory, of 1000 patients among whom 500 have pyrazinamide resistance, 450 would be detected (true positive) by tNGS and 50 would not be detected by tNGS (false negative); whereas 450 without pyrazinamide resistance would be negative by tNGS (true positive) and 50 would be reported as pyrazinamide resistant by tNGS (false positive).

Bedaquiline – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 68% (43-93) and 97% (94-100) respectively, based on up to 4 studies and up to 519 samples and a 6% prevalence of bedaquiline resistance. The indeterminate rate was 17%. Quality of evidence was judged to be low for sensitivity and high for specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have bedaquiline resistance, 20 would be detected (true positive) by tNGS and 10 would not be detected by tNGS (false negative); whereas 941 without bedaquiline resistance would be negative by tNGS (true positive) and 29 would be reported as bedaquiline resistant by tNGS (false positive).

Linezolid – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 69% (39-99) and 100% (100-100) respectively, based on up to 6 studies and up to 1093 samples and a 3% prevalence of linezolid resistance. The indeterminate rate was 15%. Quality of evidence was judged to be low for sensitivity and high for specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have linezolid resistance, 21 would be detected (true positive) by tNGS and 9 would not be detected by tNGS (false negative); whereas 970 without linezolid resistance would be negative by tNGS (true positive) and 0 would be reported as linezolid resistant by tNGS (false positive).

Clofazimine – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 70% (35-100) and 96% (93-99) respectively, based on up to 6 studies and up to 789 samples and a 3% prevalence of clofazimine resistance. The indeterminate rate was 12%. Quality of evidence was judged to be low for sensitivity and high for specificity.

Results indicate that, in theory, of 1000 patients among whom 30 have clofazimine resistance, 21 would be detected (true positive) by tNGS and 9 would not be detected by tNGS (false negative); whereas 931 without clofazimine resistance would be negative by tNGS (true positive) and 39 would be reported as clofazimine resistant by tNGS (false positive).

Amikacin – pDST as reference standard.

Pooled sensitivity and specificity (95% confidence interval) were 87% (75-100) and 99% (98-100) respectively, based on up to 8 studies and up to 1003 samples and a 10% prevalence of amikacin resistance. The indeterminate rate was 18%. Quality of evidence was judged to be very low for sensitivity and moderate for specificity.

Results indicate that, in theory, of 1000 patients among whom 100 have amikacin resistance, 87 would be detected (true positive) by tNGS and 13 would not be detected by tNGS (false negative); whereas 891 without amikacin resistance would be negative by tNGS (true positive) and 9 would be reported as amikacin resistant by tNGS (false positive).

Ethambutol – composite as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 97% (95-100) and 98% (96-100) respectively, based on 4 studies and up to 431 samples and a 78% prevalence of ethambutol resistance. The indeterminate rate was 21%. Quality of evidence was judged to be moderate for sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 300 have ethambutol resistance, 291 would be detected (true positive) by tNGS and 9 would not be detected by tNGS (false negative); whereas 686 without ethambutol resistance would be negative by tNGS (true positive) and 14 would be reported as ethambutol resistant by tNGS (false positive).

Ethambutol – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 91% (85-97) and 92% (88-96) respectively, based on 1 study and up to 213 samples and a 29% prevalence of ethambutol resistance. The indeterminate rate was 0%. Quality of evidence was judged to be low for sensitivity and specificity.

Results indicate that, in theory, of 1000 patients among whom 300 have ethambutol resistance, 273 would be detected (true positive) by tNGS and 27 would not be detected by tNGS (false negative); whereas 644 without ethambutol resistance would be negative by tNGS (true positive) and 56 would be reported as ethambutol resistant by tNGS (false positive). However, as phenotypic DST for ethambutol is known to miss borderline-resistant strains, using the composite reference standard for this drug is preferred.

Streptomycin – pDST as reference standard.
Pooled sensitivity and specificity (95% confidence interval) were 98% (96-100) and 75% (59-91) respectively, based on 5 studies and up to 493 samples and a 66% prevalence of streptomycin resistance. The indeterminate rate was 19%. Quality of evidence was judged to be high for sensitivity and low for specificity.

Results indicate that, in theory, of 1000 patients among whom 300 have streptomycin resistance, 294 would be detected (true positive) by tNGS and 6 would not be detected by tNGS (false negative); whereas 525 without streptomycin resistance would be negative by tNGS (true positive) and 175 would be reported as streptomycin resistant by tNGS (false positive).

Sub-analyses on samples from patients living with HIV or for samples with very low or low semi-quantitative results from MTB/RIF Xpert® or Xpert Ultra® were similar to the overall results. Although no data were available from children, the low/very low semi-quantitative results provide some indirect evidence of diagnostic accuracy in paucibacillary disease, which is typical in the paediatric patient population. Indeterminate rates ranged between around 10-20%, with the exception of ethambutol (pDST) which had zero indeterminate results. The data contributing to this result were only from TBSeq.

Authors' conclusions:

The impact of tNGS platforms will depend on where they are placed. In this review the quality of evidence supporting their use as an initial test of drug resistance is lower than that supporting their

use as a reflex test for patients already diagnosed with rifampicin resistance. This is largely because the datasets were strongly enriched for rifampicin resistance. Although the prevalence of resistance should not unduly impact sensitivity or specificity, the results for PICO 1 should still be interpreted with caution. Whether tNGS is placed as an initial test of drug resistance will also be influenced by its value added compared to MTB/RIF Xpert® or Xpert Ultra®. No head-to-head analysis addressing this question is reported here but future studies should explore, ideally including assessing the impact on patient-important outcomes and the value of upfront detection of isoniazid and fluoroquinolone resistance. The evidence presented in this review does however show accuracy that is comparable or better than current WHO recommended initial and follow-on tests for resistance detection, with the added value of assaying almost all drugs of interest simultaneously. Future studies will be needed to examine the diagnostic accuracy for use on extra-pulmonary specimens.

Background

Rapid, accurate detection of drug resistance for patients with tuberculosis is an important component within the process of ensuring optimal treatment outcomes both at an individual patient and population level. While empirical treatment with first line anti-tuberculosis drugs (isoniazid, rifampicin, ethambutol and pyrazinamide) would still result in successful outcomes for most patients in the world, this approach has become increasingly hazardous in some countries over the past decades.(1) Offering patients treatment regimens containing too few efficacious drugs risks treatment failure or relapse for the individual as well as amplification of resistance to the remaining drugs. At a population level this steadily selects for resistant strains with ever more patients suffering worse outcomes.(2)

Although WHO has been calling for universal access to drug susceptibility testing (DST) since the publication of the End TB Strategy in 2014,(3) the historical gold standard of phenotypic DST (pDST) has remained out of reach for most patients. The MTB/RIF Xpert® platform has made a substantial contribution to the identification of rifampicin resistance, with line-probe assays further offering characterization of resistance to fluoroquinolones and injectables where available and required. More recently the Xpert XDR® cartridge has been recommended as a reflex test to detect resistance to isoniazid, fluoroquinolones, injectables and ethionamide among rifampicin resistant strains.(4) The advantages of these molecular assays include relative simplicity of use and the potential for rapid turnaround times as they test clinical samples without the need for mycobacterial culture, and hence bio-safety level 3 (BSL-3) laboratory facilities.

After decades of little movement in the treatment of tuberculosis, new treatment options have started to emerge over the past 10 years.(5) Although WHO guidelines for treatment of rifampicin susceptible tuberculosis have not changed, there is now evidence supporting shorter treatment regimens based around rifamycins and fluoroquinolones, although it should be noted that some parts of the world have concerning levels of mono-resistance to the latter.(6) Much has changed regarding treatment of rifampicin-resistant tuberculosis, where new and re-purposed drugs are now recommended in place of injectables and other legacy drugs. Bedaquiline, pretomanid and linezolid, with or without moxifloxacin, is the newest recommended regimen for the treatment of rifampicin-resistant tuberculosis (BPaLM).(7) The pipeline of newer chemical entities undergoing trials is more richly stocked than for many decades.(8) Nevertheless, the success in drug development and regimen design have outpaced those in the necessary companion diagnostics.(9)

Sequencing of mycobacterial DNA has emerged as a promising approach to DST. With the sequencing of the great majority of genomic targets within reach and a detailed enough interpretative knowledgebase to transform sequence data into DST predictions, some countries have already moved to implement whole-genome sequencing (WGS) as the primary diagnostic assay of choice.(10,11) However, WGS remains unreliable without a prior culture step and therefore does not constitute a solution where rapid results are required and where BSL-3 laboratory facilities remain unavailable to handle mycobacterial cultures.(12) So-called targeted Next Generation Sequencing (tNGS) solutions have emerged as a potential solution. tNGS amplifies relevant targets within the genome of bacteria in the primary clinical sample before sequencing the amplicons on modern genome sequencing platforms. This yields great sequencing depth at relevant genomic loci to inform drug resistance predictions without the need for prior mycobacterial culture.

A key potential advantage of tNGS platforms over other direct-from-sample molecular assays is the inclusion of targets relevant to new and re-purposed drugs. Although none yet provide predictions for pretomanid, tNGS platforms do make predictions for the other drugs within the BPaLM regimen. As resistance to bedaquiline, linezolid and moxifloxacin is already circulating, the empirical use of this regimen risks the amplification of resistance to the component drugs.(9) Rapid and accurate diagnostics aim to maximize the benefits of such novel regimens.

Clinical pathway

The analyses presented here first assess the diagnostic accuracy of tNGS assays for DST for isoniazid, rifampicin, ethambutol, pyrazinamide and fluoroquinolones (levofloxacin and moxifloxacin) among all patients diagnosed with TB and then assess the diagnostic accuracy for all covered drugs (except rifampicin) for patients whose strains have already been identified as resistant to rifampicin by Xpert MTB/RIF® or Xpert Ultra®. Further sub-group analyses examine the performance by semi-quantitative Xpert MTB/RIF® or Xpert Ultra® result ('very low' or 'low' compared to 'medium' or 'high' bacterial load based on cycle threshold value), and for patients living with HIV infection, as defined by testing as part of an included study. No data were available on the performance of these assays in children, but the semi-quantitative results can be interpretated as a proxy for paucibacillary disease commonly seen in the paediatric population. All analyses are restricted to respiratory samples.

Historically, molecular assays have reported one of two outcomes (besides assay failure): either 'Resistant' upon detecting a resistance conferring mutation, or 'Not resistant' where no such mutation has been detected (from which 'susceptible' is usually inferred). This has led to each assay having one of four potential outcomes:

True positives: the assay correctly predicts resistance according to the reference standard. Patient correctly treated with appropriately modified regimen for resistance pattern. Risk of treatment failure or developing further resistance minimized.

False positive: the assay predicts resistance but the reference standard reports susceptibility, with the consequence that a patient might not receive a drug they could potentially benefit from. Patient incorrectly shifted to a more aggressive treatment regimen, increasing risk of adverse effects unnecessarily, loss to follow-up, and emergence of further resistance.

True negative: the assay reports that no resistance has been detected and the strain is indeed susceptible according to the reference standard. Patient correctly treated with appropriate regimen. Treatment burden minimized.

False negative: the assay reports that no resistance has been detected but resistance has been detected by the reference standard. In this case, a patient might receive an inefficacious drug with the risk that they suffer a poor treatment outcome and relapse with additional resistance to other drugs (what is known as 'amplification of resistance'). Patient incorrectly treated with inappropriate regimen, increasing risk of treatment failure, mortality, amplification of resistance, and transmission of drug resistant TB.

In addition to the two standard outcomes reported by most other molecular assays to date, Genoscreen's Deeplex® Myc-TB reports 'non-synonymous uncharacterized variant or uncharacterized indel' when it encounters a mutation within a target sequence that is not listed as predictive of resistance or as consistent with susceptibility in its interpretative catalogue. A prediction of 'S' by Deeplex® Myc-TB is therefore no longer an inference based on the absence of a detected resistance mutation, but an active call that asserts that all identified bases in the target sequence are consistent with susceptibility.

For the evaluation of diagnostic test accuracy, it is customary to assess sensitivity and specificity based only upon the assay's determinate results ('R' or 'S'), and not to include indeterminate results in the denominator. These can take one of three forms: 'failed target amplification', which is relevant to particular drugs within a sample; 'failed sample'; and in the case of Deeplex® Myc-TB, the presence of an uncharacterized mutation in the absence of a resistance mutation. This analysis follows this precedent, but will report the number and proportion of indeterminate results as well, where these are a composite of 'failed sample', 'failed target amplification', and for Deeplex® Myc-TB 'the presence of an uncharacterized mutation in the absence of a resistance mutation'.

PICOs

Generically stated, this review explored two PICOs. Each is applied to a specific list of drugs:

1. Should tNGS as the initial test be used to diagnose drug resistance in patients with bacteriologically confirmed pulmonary TB disease?

   Applies to
   - Rifampicin, using a composite reference standard of pDST and WGS and MTB/RIF Xpert® or Xpert Ultra®
   - Rifampicin, using pDST as the reference standard
   - Isoniazid, using pDST as the reference standard
   - Levofloxacin, using pDST as the reference standard
   - Moxifloxacin, using pDST as the reference standard
   - Pyrazinamide, using a composite reference standard of pDST and WGS
   - Pyrazinamide, using only pDST as the reference standard
   - Ethambutol, using a composite reference standard of pDST and WGS
   - Ethambutol, using only pDST as the reference standard

2. Should tNGS be used to diagnose drug resistance in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?
   Applies to
   - Isoniazid, using pDST as the reference standard
   - Levofloxacin, using pDST as the reference standard
   - Moxifloxacin, using pDST as the reference standard
   - Pyrazinamide, using a composite reference standard of pDST and WGS
   - Pyrazinamide, using only pDST as the reference standard
   - Bedaquiline, using pDST as the reference standard
   - Linezolid, using pDST as the reference standard
   - Clofazimine, using pDST as the reference standard
   - Amikacin, using pDST as the reference standard
   - Ethambutol, using a composite reference standard of pDST and WGS
   - Ethambutol, using only pDST as the reference standard
   - Streptomycin, using pDST as the reference standard

Sub-analyses were performed to explore diagnostic test accuracy in patients living with HIV, and for semi-quantitative results (derived from cycle thresholds) from MTB/RIF Xpert® or Xpert Ultra®, where ‘very low’ or ‘low’ concentrations of *M. tuberculosis* were compared to ‘medium’ or ‘high’ concentrations. As no data on children were available, the semi-quantitative results can be interpreted as a proxy for paucibacillary disease.

Review objective

To assess the diagnostic accuracy of tNGS platforms positioned as either the initial test to be used to diagnose drug resistance for patients with bacteriologically confirmed pulmonary TB, or positioned as a reflex test for patients with bacteriologically confirmed rifampicin-resistant pulmonary TB.

Methods

Types of studies and participants

Data were included from both published and unpublished, prospective, observational clinical studies of tNGS platform diagnostic accuracy. We included all studies where tNGS had been performed directly from clinical samples and excluded studies where tNGS had been performed exclusively on cultured isolates. We included all samples as long as data on tNGS result and reference had been finalized and were not subject to change. All studies were required to have comparator pDST data as a reference, and for rifampicin, ethambutol and pyrazinamide were required to also have WGS to allow a composite reference to be generated. Rifampicin resistance results and semi-quantitative results from MTB/RIF Xpert® or Xpert Ultra® were requested from all studies.

Drugs and reference standards

We assessed diagnostic accuracy for:

Isoniazid	-	pDST
Rifampicin	-	pDST + WGS + MTB/RIF Xpert® or Xpert Ultra® (composite)
Ethambutol	-	pDST + WGS (composite)
Pyrazinamide	-	pDST + WGS (composite)
Moxifloxacin	-	pDST
Levofloxacin	-	pDST
Amikacin	-	pDST
Kanamycin	-	pDST
Capreomycin	-	pDST
Streptomycin	-	pDST
Linezolid	-	pDST
Clofazimine	-	pDST
Bedaquiline	-	pDST

The composite reference was considered ‘R’ if either pDST or WGS, or in the case of rifampicin, MTB/RIF Xpert® or Xpert Ultra®, were ‘R’, and the composite was considered ‘S’ if both pDST and WGS, and in the case of rifampicin, MTB/RIF Xpert® or Xpert Ultra®, were ‘S’. If results for one of pDST or WGS, or MTB/RIF Xpert® or Xpert Ultra®, were missing, then no composite result was generated.

Search methods

An Oxford University librarian (report co-author Eli Harriss) used the following search terms to search Ovid Medline, Ovid Embase and Scopus on 7th September 2022:

(("Tuberculosis"[Mesh]) OR (tuberculosis[Text Word] OR TB[Text Word])) AND (("High-Throughput Nucleotide Sequencing"[Mesh]) OR ("next generation sequencing"[Text Word] OR "deep sequencing"[Text Word] OR tNGS[Text Word] OR "targeted sequencing"[Text Word] OR "amplicon sequencing"[Text Word] OR Deepcheck[Text Word] OR Deeplex*[Text Word] OR NanoTB[Text Word]))

The search was repeated on January 17th 2023 using the same search terms.

We used the Deduklick programme to reliably remove the duplicated results. Two reviewers (Timothy Walker and Phu Phan Trieu) used the web-based programme Rayyan to independently reviewed the titles, and then abstracts and full texts where indicated.

In addition, to a WHO public call for data, we also contacted well known experts in the field to ask if they have, or know of, unpublished data that could be contributed.

Inclusion criteria were set as follows, and all needed to be met:
- Studies using 'design locked' and 'market ready' tNGS platforms
- Studies assessing samples from patients diagnosed with bacteriologically confirmed tuberculosis disease by culture, WHO recommended molecular assay or lateral flow lipoarabinomannan (LF-LAM)
- Studies using a gold standard comparator of phenotypic drug susceptibility testing (pDST) or a composite of culture based whole genome sequencing genomic DST (gDST) for rifampicin, ethambutol and pyrazinamide.
- Studies reporting tNGS based gDST results for any of the following drugs: isoniazid, rifampicin, moxifloxacin, levofloxacin, bedaquiline, pyrazinamide, linezolid, pretomanid, delamanid, clofazimine, amikacin, ethambutol, ethionamide, prothionamide, streptomycin.
- Studies for which individual patient data were available, including after correspondence with the authors (see justification in statistics section below).

Exclusion criteria
- Data not yet finalized or still subject to change

Data extraction

As few studies were identified through the literature search, we included all data identified from those studies after correspondence with the authors. As such, no manual data extraction from the publication was required. We made a post-hoc decision to perform only an individual patient data (IPD) meta-analysis (see statistics section below). For this reason, we excluded any study that could not provide individual patient data.

Assessment of methodological quality

Two report authors made independent assessments of methodological quality using QUADAS-2.(13) Disagreements were resolved by discussion and uncertainties or disagreements reviewed by an independent third party.

QUADAS-2 domains along with relevant signalling questions are detailed below. In each case, the answer was 'unclear' where the relevant information was unavailable. Answers were otherwise stated as 'low risk' or 'high risk', in keeping with the QUADAS-2 terminology.

<u>Domain 1: Patient selection</u>

A. Risk of bias: could patient selection have introduced bias?
    i. Signalling question 1: was a consecutive or random sample of patients/specimens enrolled?
    We scored 'high risk' if the study selected patients by convenience.
    ii. Signalling question 2: was the study design prospective?
    We scored 'high risk' if the study was retrospective
    iii. Signalling question 3: did the study avoid inappropriate exclusions?
    We scored 'high risk' if the study excluded patients with low risk of rifampicin resistance

B. Applicability: Are there concerns that the included patients and setting do not match the review question?
    We scored 'high risk' for PICO 1 where the samples were enriched for rifampicin resistance.

<u>Domain 2: Index test</u>

A. Risk of bias: could the conduct or interpretation of the index test have introduced bias?
   i. Signalling question 1: was the index test result interpreted without knowledge of the reference test result?
   We scored 'high risk' if the operator of the index test was unblinded to the result of the reference. Although the output of the index test is automated, there is still some room for interpretation by the operator, such as when the call is based on low frequency resistance alleles. We also observed that some operators initially wrongly interpreted the meaning of Deeplex® Myc-TB's 'non-synonymous uncharacterized variant or uncharacterized indel' as 'R' in some cases and as 'S' in others, instead of 'U' (indeterminate) as intended by the assay. Although we corrected these errors of interpretation, they indicate room for operator bias.
   ii. Signalling question 2: if a threshold was used, was it prespecified?
   We scored 'low risk' for all as thresholds for interpretation are defined by the assay manufacturers.

B. Applicability: Are there concerns that the index test, its conduct, or its interpretation differ from the review question?
   We scored 'low risk' for all on the basis that the index tests were performed as recommended by the manufacturers.

Domain 3: Reference standard

A. Risk of bias: Could the reference standard, its conduct, or its interpretation have introduced bias?
   i. Signalling question 1: Is the reference standard likely to correctly classify the target condition?
   We scored 'high risk' where a non WHO-recommended critical concentration was used for pDST.
   ii. Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?
   We scored 'high risk' if the operator was unblinded to the result of the index test.
   iii. Signalling question 3: Were the index test and reference standard done on the same type of sample?
   We scored 'high risk' if different samples were used.

B. Applicability: Are there concerns that the target condition as defined by the reference standard does not match the question?
   We scored 'low risk' for all studies.

Domain 4: Flow and timing

A. Risk of Bias: Could the patient flow have introduced bias?
   i. Signalling question 1: Was there an appropriate interval between the index test and reference standard?
   We scored 'low risk' wherever the same sample was used for the index and reference test. Where this was not the case and where samples were separated by more than a week in time, we scored 'high risk'.
   ii. Signalling question 2: Did all patients receive the same reference standard?
   We scored 'low risk' for all as the reference standards were consistent within studies
   iii. Signalling question 3: Were all patients included in the analysis?
   We scored 'low risk' for all as we were in contact with each data contributor to make sure that all the available data were included.

We assessed the presented evidence using the GRADE (Grading of Recommendation Assessment, Development and Evaluation) methodology. Certainty of evidence was graded as:
- High (not downgraded)

- Moderate (downgraded 1 level)
- Low (downgraded 2 levels)
- Very low (downgraded by >2 levels)

based on

- Risk of bias
- Indirectness
- Inconsistency
- Imprecision
- Publication bias

We did not downgrade the risk of bias where QUADAS-2 scored 'high risk' for PICO 1 because of studies enriching for rifampicin resistant samples. Instead, we downgraded one for indirectness where this was the case. We did not downgrade for risk of bias because the prevalence of resistance should not have an undue impact on sensitivity and specificity, unlike for positive and negative predictive values. Where a non WHO-recommended critical concentration had been used, we downgraded one for risk of bias. We downgraded one for inconsistency where there was a study that was clearly an outlier in performance as judged by the forest plots that summarize the raw data that were used for the models. We downgraded one for imprecision where the 95% confidence intervals around the point estimates for pooled sensitivity were greater than 20 percentage points and for pooled specificity were greater than 10 percentage points. Further details of particular judgements are provided in the results section.

Statistical analysis and data synthesis

Outcomes were described according to the tNGS result and the reference:

Where the reference was R:

TP = true positive
FN = false negative

Where the reference was S:

TN = true negative
FP = false positive

The number and percentage of indeterminate results was recorded, where this was a composite of 'failed sample', 'failed target amplification', and for Deeplex® Myc-TB, 'the presence of an uncharacterized mutation in the absence of a resistance mutation', otherwise known as a 'U' (unknown) mutation.

We first reviewed the pooled data and excluded drugs for which fewer than 50 samples were resistant according to the reference test, or for which fewer than 100 samples were susceptible. We made an exception for new or re-purposed drugs (Bedaquiline, Clofazimine, Linezolid) as these are of particular interest and as we expected there to be few resistant samples available.

As this is a review of the diagnostic accuracy of a class of diagnostic platforms, and not any particular product, we next analysed all the data from each platform alone to assess which to include in an analysis to inform a class recommendation and to benchmark acceptable performance in the technical class going forward. Where the performance of any one platform appeared as an outlier for sensitivity or specificity, that platform was excluded from subsequent meta-analyses. We judged a platform to be an outlier for a particular drug if the point estimate for sensitivity was more than 10 percentage points worse than the best performing platform, or more than 5 percentage points worse for specificity. We also set a minimum number of susceptible samples (50) and a minimum number of resistant samples (25) per platform per drug. With the exclusion of Bedaquiline, Clofazimine, and Linezolid, results for platforms were excluded where insufficient data on a drug were available. These *a priori* rules were intended to allow for a

contingent class recommendation to be made, setting minimal performance characteristics which individual platforms within class will have to meet in the future.

We performed and individual patient data (IPD) meta-analysis instead of a classical meta-analysis. There were a number of reasons for this:

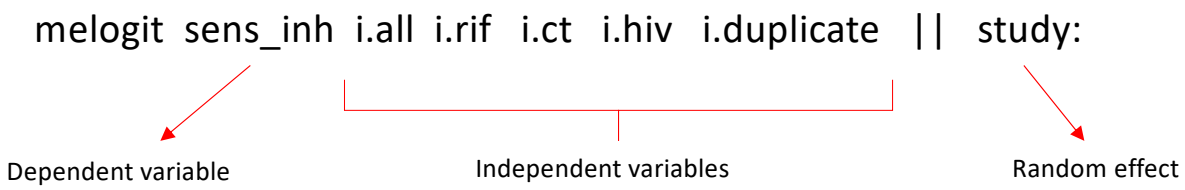
- The studies identified in the literature were generally too small to contribute to a classical meta-analysis
- An IPD meta-analysis allowed for co-variables to be included in the model
- A large proportion of the data came from FIND where the same samples were sequenced on different platforms. The IPD approach allowed us to control for this.

For each of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), we generated a binary dependent variable (1/0):

- Sensitivity=1 if true positive; Sensitivity=0 if false negative
- Specificity=1 if true negative; Specificity =0 if false positive
- NPV=1 if true negative; NPV=0 if false negative
- PPV=1 if true positive; PPV=0 if false positive

For each dependent variable we then built a multivariable model in which we included a number of co-variables as fixed effects. These included rifampicin resistance as determined by MTB/RIF Xpert® or Xpert Ultra® for all drugs other than rifampicin; semi-quantitative CT value from MTB/RIF Xpert® or Xpert Ultra®; and a co-variable to indicate which samples featured in duplicate and which not (all FIND samples were assayed once by Deeplex® Myc-TB and once by NanoTB®). We initially included this variable as a random effect but the model failed to converge, so it was included as a fixed effect. For models looking specifically at diagnostic test accuracy in HIV infection, we also included HIV test results as a co-variable. However, as data on HIV were only available from FIND, we did not include HIV in the main models to avoid losing all the data from other studies. Finally, we included the study site as a random effect.

The models were run in STATA (version 17) using the melogit command, and the outputs transformed using the margins command:



In this example the dependent variable is ‘sensitivity for isoniazid’. We include an additional variable ‘all’ that is coded uniformly as ‘1’ in order to capture the overall effect with the margins command.

We ran models for all PICO questions for sensitivity, specificity, NPV and PPV.

No data on patient outcomes were available.

Using MTB/RIF Xpert® or Xpert Ultra® results

Data on rifampicin resistance by MTB/RIF Xpert® or Xpert Ultra® were used for two different purposes. First, these results formed part of the composite reference standard for rifampicin. In this case it was important to use the MTB/RIF Xpert® or Xpert Ultra® obtained from the same sample that was used for the index test. Second, these results informed the co-variable ‘rifampicin resistance’ in the multivariable models.

Whilst each study that provided us with data on MTB/RIF Xpert® or Xpert Ultra® did so for just a single result, the FIND data included results on two MTB/RIF Xpert® or Xpert Ultra® results. The first was from the screening test for entry into the study and the second was repeated as part of the study on the same sample that was also processed for the index test. We used the repeat test results from FIND to inform the composite reference standard for rifampicin, and the semi-quantitative results. However, we used the result of the screening test for the co-variable in the multi-variable model. We judged that this result best represented how tNGS might be positioned in the real world in the future: as a reflex test after MTB/RIF Xpert® or Xpert Ultra®. We did not consider it likely that an MTB/RIF Xpert® or Xpert Ultra® would be repeated alongside tNGS in the real world after an initial test on another sample had already been performed.

Results

Results of the literature search and call for data

We identified 876 articles from the literature review. Nine full texts were reviewed after 867 articles were excluded on the basis of title alone. Three met inclusion criteria. An additional published article was identified after contacting experts in the field.

Sources of published data:

The four papers identified from the published literature were:

1. Feuerriegel *et. al.* "Rapid genomic first – and second-line drug resistance prediction from clinical *Mycobacterium tuberculosis* specimens using Deeplex-MycTB", European Respiratory Journal 2021 Jan 5;57(1):2001796.

This study used a convenience sampling frame from a supra-national reference laboratory in Germany. Sputum samples were frozen, then decontaminated using 3% NaOH/NALC and heat inactivated.

2. Kambli *et. al.* "Targeted next generation sequencing directly from sputum for comprehensive genetic information on drug resistant Mycobacterium tuberculosis", Tuberculosis (Edinb). 2021 Mar;127:102051.

This study operated from an MDR-TB referral centre in Mumbai, India, and further selected for patients with a high risk of MDR-TB. Sputum samples were decontaminated using NaLC/NaOH, 1% final NaOH concentration, and sediments resuspended in phosphate buffer.

3. Bonnet *et. al.*, "A Comprehensive Evaluation of GeneLEAD VIII DNA Platform Combined to Deeplex Myc-TB® Assay to Detect in 8 Days Drug Resistance to 13 Antituberculous Drugs and Transmission of Mycobacterium tuberculosis Complex Directly From Clinical Samples", Front. Cell. Infect. Microbiol. 11:707244

This study operated out of the French national referral centre, Paris, and selected consecutive samples referred for evaluation of drug resistant TB. Sputum samples were stored overnight at 4°C when they could not be processed on the day of receipt.

4. Mansoor *et. al.,* "Clinical utility of target-based next-generation sequencing for drug-resistant TB", International Journal of Tuberculosis and Lung Disease 2023, 27(1):41–48

This study started out recruiting consecutive samples in Mumbai, India, and then switched to only selecting MDR-TB half way through to obtain more drug resistant samples.

Sources of unpublished data:

The single largest study was contributed by FIND from recently completed work running head-to-head comparisons of Deeplex® Myc-TB and NanoTB® on samples in three settings: South Africa, India and Georgia. This study selected patients at risk of rifampicin resistance and tested each sample on both tNGS platforms. Relevant inclusion criteria: Adults with *M. tuberculosis* detected by Xpert MTB/RIF® or Xpert Ultra®, AND rifampicin resistance detected by the same assay OR sputum smear or culture positive after 3 months of standard TB treatment OR previously diagnosed rifampicin resistant TB or sputum or culture positive after 3 months of standard MDR-TB therapy OR previously received more than a month of treatment for a prior TB episode OR close contact of a patient with known drug-resistant TB. This study used decontaminated sputum samples that were frozen until a sufficient number of samples for a tNGS sequencing run had

been collected.

The “DIAgnostics for Multidrug Resistant Tuberculosis in Africa” (DIAMA) study is on-going and set in Rwanda, Cameroon, Ethiopia, Guinea, Mali, and Benin. It is has enriched for patients with rifampicin resistance, aiming to recruit approximately an equal number of patients with and without rifampicin resistance. It is assessing the diagnostic accuracy of Deeplex® Myc-TB.

The icddr,b in Bangladesh contributed data from an on-going study of Deeplex® Myc-TB in Bangladesh. Samples have been collected in an unselected manner.

The National Institute for Communicable Diseases in South Africa contributed data from an evaluation of Deeplex® Myc-TB as applied to samples diagnosed as rifampicin resistant by Xpert MTB/RIF®, and samples that were rifampicin susceptible by Xpert MTB/RIF® and collected as part of a surveillance programme. Results were from remnant decontaminated sputum samples digested by NALC-NaOH in preparation for Xpert MTB/RIF® or Xpert Ultra® testing.

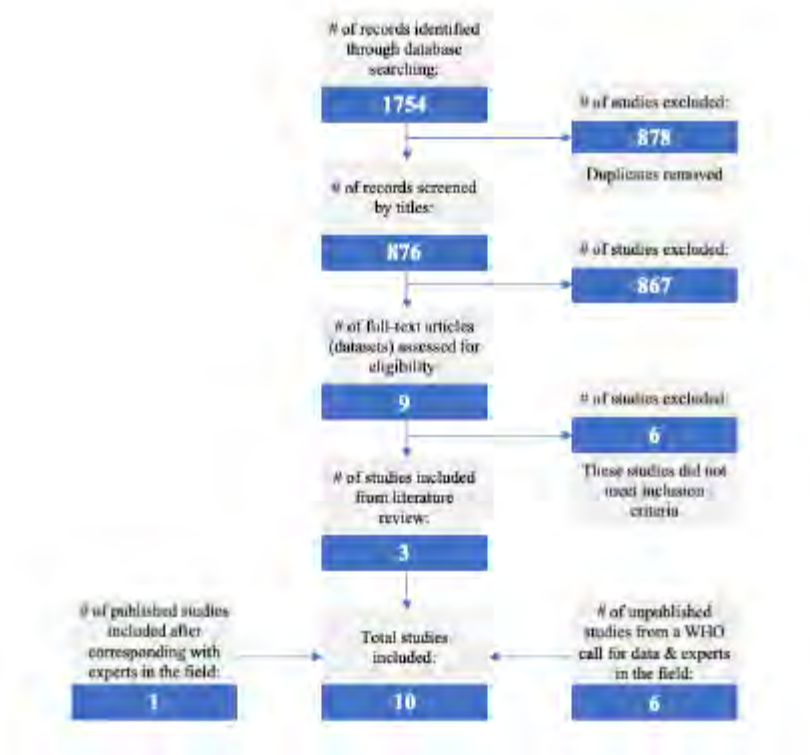
The San Raffaele Scientific Institute (SRSI) in Italy contributed a small number of samples from Italy and Eritrea. This was a convenience sample obtained through the supra-national reference laboratory’s activities.

An on-going collaboration between ShengTing Biotech and the Beijing Chest Hospital resulted in the contribution of data from the TBseq® (ShengTing Biotech) platform. These data have been shared by the manufacturer but the study was run independently by the Beijing Chest Hospital. The study enriched for patients with MDR-TB.

All subjects from all studies were adults, to the best of our knowledge. Data on HIV test results were available from FIND only. Data on rifampicin resistance as determined by MTB/RIF Xpert® or Xpert Ultra® were available from FIND; DIAMA; icddr,b; Kambli *et al;* Mansoor *et al*; and TBSeq. In addition to these, data on semi-quantitative CT value from MTB/RIF Xpert® or Xpert Ultra® were available from SISR.

Where not stated, it was unclear how sputum samples were processed in a study.

Flow chart 1: Flow of studies in the review



Identifying which drugs and platforms to include in the class-based assessment

There were sufficient samples among the pooled data to address each PICO. No drugs were therefore dropped from the analysis.

For each drug, between 0 and 2 platforms were dropped from the analysis based on either the overall number of resistant or susceptible samples available for that platform and drug, or based on performance characteristics that were insufficiently close to the diagnostic test accuracy of the best performing platform. Table 1 shows the relative accuracy of each platform for which we had data for each drug, as well as the overall number of resistant and susceptible samples that were available. Platforms highlighted in red were excluded. Sensitivity and specificity in table 1 are based on the raw data (the actual number of TP, FN, TN, and FP results), and are not derived from a model.

Table 1: Results of selection process of platforms for meta-analysis

Platform	Drug	R	S	Sensitivity	Specificity
Deeplex	Isoniazid	1056	498	97.0	95.6
NanoTB	Isoniazid	505	144	95.4	98.6
TBSeq	Isoniazid	229	221	90.4	79.6
Deeplex	Rifampicin	1083	548	99.1	89.1
NanoTB	Rifampicin	497	152	97.4	82.2
TBSeq	Rifampicin	285	165	94.4	66.7
Deeplex	Rifampicin_composite	915	290	95.9	96.9
NanoTB	Rifampicin_composite	534	115	95.7	100.0
TBSeq	Rifampicin_composite	78	0	100.0	n/a
Deeplex	Ethambutol	465	444	95.1	81.3
NanoTB	Ethambutol	383	286	84.9	80.1
TBSeq	Ethambutol	104	346	88.5	93.9
Deeplex	Ethambutol_composite	447	225	96.0	99.1
NanoTB	Ethambutol_composite	445	224	85.4	99.1
TBSeq	Ethambutol_composite	38	57	97.4	98.2
Deeplex	Pyrazinamide	443	425	85.8	93.9
NanoTB	Pyrazinamide	344	256	75.0	96.5
TBSeq	Pyrazinamide	0	17	0.0	5.9
Deeplex	Pyrazinamide_composite	397	269	89.2	98.5
NanoTB	Pyrazinamide_composite	358	242	74.6	100.0
Deeplex	Moxifloxacin	371	640	97.3	95.6
NanoTB	Moxifloxacin	302	367	94.7	97.5
Deeplex	Levofloxacin	354	556	97.2	96.0
NanoTB	Levofloxacin	303	365	93.7	97.0
TBSeq	Levofloxacin	107	343	82.2	87.8
Deeplex	Amikacin	65	806	92.3	99.4
NanoTB	Amikacin	58	622	87.9	99.2
Deeplex	Streptomycin	511	264	97.8	88.3
NanoTB	Streptomycin	440	203	81.4	89.7
TBSeq	Streptomycin	116	333	83.6	84.4
Deeplex	Bedaquiline	40	771	85.0	97.5
NanoTB	Bedaquiline	34	604	5.9	99.7
Deeplex	Clofazimine	36	794	80.6	97.5
NanoTB	Clofazimine	32	606	0.0	100.0
Deeplex	Linezolid	32	810	46.9	99.9
NanoTB	Linezolid	31	638	48.4	99.8
TBSeq	Linezolid	1	0	100.0	0.0

R=”Resistant” by reference standard”; S=”Susceptible by reference standard”. Sensitivity and specificity indicate a percentage.

After exclusion of the platforms highlighted in red in table 1, data were available for the following drugs from the following platforms (table 2) for inclusion in the IPD meta-analysis. The colours in table 2 are to differentiate the platforms and have no other significance. The number of samples per study are plotted in figure 1.

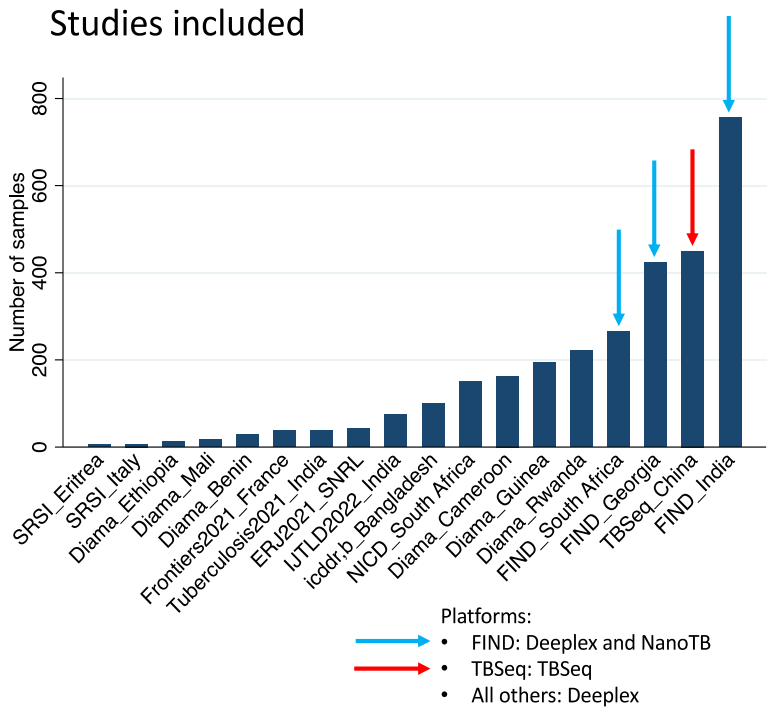
Table 2: Platforms and drugs included in the IPD meta-analysis

Drug	Deeplex	NanoTB	TBSeq
Isoniazid	1554	649	
Rifampicin_composite	1205	649	
Rifampicin	1631		
Ethambutol_composite	672		95
Ethambutol			450
Pyrazinamide_composite	666		
Pyrazinamide	868		
Levofloxacin	910	668	
Moxifloxacin	1011	669	
Streptomycin	775		
Amikacin	871	680	
Kanamycin	1334	617	
Capreomycin	1210	639	
Bedaquiline	811		
Linezolid	842	669	
Clofazimine	830		

Figure 1: Sources of data identified

From the literature:
- ERJ 2021 (SNRL Germany)
- Frontiers 2021 (France)
- Tuberculosis 2021 (India)
- IJTLD 2022 (India)

Unpublished:
- FIND
  - Georgia
  - India
  - South Africa
- Diama
  - Benin
  - Guinea
  - Cameroon
  - Rwanda
  - Mali
  - Ethiopia
- NICD, South Africa
- icddr,b (Bangladesh)
- San Raffaele Scientific Institute, Italy
  - Italy
  - Eritrea
- TBSeq, China



The results of the IPD meta-analysis showed a sensitivity was over 85% for all drugs with the exception of Bedaquiline, Linezolid and Clofazimine for which sensitivity was 68%, 69%, and 70%, respectively. Specificity was over 90% for all drugs with the exception of Streptomycin (75%) and RIF (81%) where only the pDST was used as the reference standard (table 3).

Table 3: Summary of the results from the IPD meta-analysis addressing each PICO

		Sensitivity	Specificity	Negative predictive value	Positive predictive value
PICO 1					
1	Rifampicin_composite	93.1 (87.0 - 99.2)	96.2 (88.6 - 100)	84.8 (78.4 - 91.1)	99.0 (96.5 - 100)
2	Isoniazid	95.8 (92.8 - 98.7)	97.0 (95.1 - 98.9)	90.9 (87.7 - 94.0)	98.9 (98.2 - 99.6)
3	Levofloxacin	94.2 (88.4 - 99.9)	96.2 (93.4 - 98.9)	96.8 (95.7 - 97.9)	93.2 (88.5 - 97.9)
4	Moxifloxacin	95.6 (92.4 - 98.7)	96.3 (93.2 - 99.5)	97.4 (96.2 - 98.7)	94.4 (90.6 - 98.3)
5	Pyrazinamide_composite	88.4 (85.2 - 91.7)	98.5 (97.1 - 100)	87.7 (80.6 - 94.9)	98.7 (97.5 - 99.9)
6	Ethambutol_composite	95.8 (94.0 - 97.6)	99.3 (98.2 - 100)	93.4 (89.3 - 97.4)	99.5 (98.8 - 100)
7	Ethambutol	88.0 (81.7 - 94.3)	94.0 (91.5 - 96.6)	96.1 (93.4 - 98.7)	82.7 (75.2 - 90.1)
8	Pyrazinamide	85.3 (80.2 - 90.4)	93.9 (91.6 - 96.3)	83.5 (74.6 - 92.5)	92.4 (87.6 - 97.2)
9	Rifampicin	98.7 (97.2 - 100)	81.0 (69.5 - 92.5)	97.6 (96.1 - 99.2)	94.4 (92.4 - 96.3)
PICO 2					
10	Isoniazid	96.5 (93.8 - 99.2)	95.8 (91.8 - 99.8)	75.7 (67.1 - 84.4)	99.6 (99.1 - 100)
11	Levofloxacin	95.8 (90.4 - 100)	96.0 (93.1 - 98.9)	97.5 (96.2 - 98.8)	94.2 (89.6 - 98.9)
12	Moxifloxacin	96.5 (93.6 - 99.5)	95.2 (91 - 99.4)	97.4 (95.6 - 99.2)	94.7 (90.9 - 98.6)
13	Pyrazinamide_composite	90.0 (86.8 - 93.2)	98.6 (96.8 - 100)	84.6 (75.2 - 94.0)	99.3 (98.4 - 100)
14	Bedaquiline	67.9 (42.6 - 93.2)	97.0 (94.3 - 99.7)	99.4 (98.6 - 100)	62.2 (46.5 - 77.8)
15	Linezolid	68.9 (38.7 - 99.1)	99.8 (99.6 - 100)	99.8 (99.4 - 100)	93.0 (84.0 - 100)
16	Clofazimine	70.4 (34.6 - 100)	96.3 (93.2 - 99.3)	99.2 (98.1 - 100)	44.2 (12.4 - 75.9)
17	Amikacin	87.4 (74.5 - 100)	99.0 (98.4 - 99.6)	98.0 (96.0 - 100)	82.0 (57.0 - 100)
18	Ethambutol_composite	96.7 (95.0 - 98.4)	98.4 (96.1 - 100)	88.8 (81.2 - 96.3)	100 (99.0 - 100)
19	Streptomycin	98.1 (96.1 - 100)	75.0 (59.5 - 90.5)	90.8 (82.0 - 99.7)	94.8 (92.8 - 96.8)
20	Ethambutol	91.0 (85.1 - 96.9)	92.0 (88.4 - 95.7)	96.0 (93.0 - 99.0)	83.0 (75.0 - 90.0)
21	Pyrazinamide	89.5 (84.5 - 94.5)	90.4 (86.4 - 94.4)	82.9 (73.1 - 92.8)	93.0 (88.3 - 97.7)

The findings of the sub-analyses by HIV test result and semi-quantitative MTB/RIF Xpert are in table 4. This shows pooled sensitivity and specificity for the output of the models along with the absolute number or TP, FN, TN and FP results (including samples that were included in the models and those not), along with the number and percentage of indeterminate results.

Table 4: Results of sub-analyses on semi-quantitative results and for patients living with HIV

Drug	Sub-group	TP	FN	Sensitivity	TN	FP	Specificity	Indeterminate (%)
Isoniazid	Semi-quantitative Xpert result: **medium or high**	1239	47	96 (93 - 99)	437	10	98 (96 - 99)	147 (8%)
Isoniazid	Semi-quantitative Xpert result: **low or very low**	149	6	97 (94 - 100)	66	5	93 (87 - 100)	136 (38%)
Isoniazid	HIV test result: **Positive**	70	18	86 (73 - 98)	48	1	*96 (89 - 100)	59 (30%)
Isoniazid	HIV test result: **Negative**	725	21	95 (91 - 100)	208	3	98 (96 - 100)	111 (10%)
Rifampicin_composite	Semi-quantitative Xpert result: **medium or high**	1221	46	93 (87 - 99)	339	5	96 (89 - 100)	73 (4%)
Rifampicin_composite	Semi-quantitative Xpert result: **low or very low**	157	12	92 (87 - 99)	40	2	93 (81 - 100)	127 (38%)
Rifampicin_composite	HIV test result: **Positive**	125	7	93 (82 - 100)	26	0		38 (19%)
Rifampicin_composite	HIV test result: **Negative**	738	29	92 (80 - 100)	200	0		101 (9%)
Rifampicin	Semi-quantitative Xpert result: **medium or high**	846	8	99 (97 - 100)	331	46	82 (71 - 93)	47 (4%)
Rifampicin	Semi-quantitative Xpert result: **low or very low**	106	1	99 (97 - 100)	39	9	72 (53 - 90)	86 (36%)
Rifampicin	HIV test result: **Positive**	59	2	95 (82 - 100)	13	8	68 (35 - 100)	16 (16%)
Rifampicin	HIV test result: **Negative**	348	3	98 (95 - 100)	113	16	70 (39 - 100)	54 (10%)
Ethambutol_composite	Semi-quantitative Xpert result: **medium or high**	363	17	96 (94 - 98)	206	1	*99 (97 - 100)	52 (8%)
Ethambutol_composite	Semi-quantitative Xpert result: **low or very low**	51	1	98 (94 - 100)	60	1	*97 (92 - 100)	62 (35%)
Ethambutol_composite	HIV test result: **Positive**	23	2	**93 (83 - 100)	52	0		21 (21%)
Ethambutol_composite	HIV test result: **Negative**	288	11	**96 (94 - 98)	148	1		86 (16%)
Ethambutol	Semi-quantitative Xpert result: **medium or high**	63	10	*96 (89 - 100)	263	16	*89 (78 - 99)	0 (0%)
Ethambutol	Semi-quantitative Xpert result: **low or very low**	25	2	*89 (81 - 97)	51	4	*93 (89 - 97)	0 (0%)
Ethambutol***	HIV test result: **Positive**	0	0		0	0		0 (0%)
Ethambutol***	HIV test result: **Negative**	0	0		0	0		0 (0%)
Pyrazinamide_composite	Semi-quantitative Xpert result: **medium or high**	283	36	89 (85 - 92)	240	3	99 (97 - 100)	39 (6%)
Pyrazinamide_composite	Semi-quantitative Xpert result: **low or very low**	23	4	86 (74 - 99)	25	1	96 (89 - 100)	69 (57%)
Pyrazinamide_composite	HIV test result: **Positive**	22	5	83 (69 - 96)	45	1	98 (94 - 100)	25 (26%)
Pyrazinamide_composite	HIV test result: **Negative**	234	28	89 (86 - 93)	194	3	98 (97 - 100)	75 (14%)

Pyrazinamide	Semi-quantitative Xpert result: **medium or high**	337	55	86 (80 - 91)	329	19	94 (92 - 97)	47 (6%)
Pyrazinamide	Semi-quantitative Xpert result: **low or very low**	27	6	82 (69 - 95)	30	5	89 (79 - 99)	71 (51%)
Pyrazinamide	HIV test result: **Positive**	21	3	89 (77 - 100)	47	2	97 (93 - 100)	25 (26%)
Pyrazinamide	HIV test result: **Negative**	224	26	89 (86 - 93)	196	13	93 (90 - 97)	75 (14%)
Levofloxacin	Semi-quantitative Xpert result: **medium or high**	573	25	94 (88 - 100)	775	27	96 (94 - 99)	41 (3%)
Levofloxacin	Semi-quantitative Xpert result: **low or very low**	52	4	95 (89 - 100)	107	5	95 (90 - 100)	102 (38%)
Levofloxacin	HIV test result: **Positive**	28	1	99 (97 - 100)	122	5	96 (91 - 100)	40 (20%)
Levofloxacin	HIV test result: **Negative**	426	26	87 (73 - 100)	507	16	95 (89 - 100)	91 (9%)
Moxifloxacin	Semi-quantitative Xpert result: **medium or high**	576	20	96 (93 - 99)	782	25	97 (94 - 100)	41 (3%)
Moxifloxacin	Semi-quantitative Xpert result: **low or very low**	52	4	93 (85 - 100)	110	5	95 (88 - 100)	104 (38%)
Moxifloxacin	HIV test result: **Positive**	30	1	99 (98 - 100)	122	3	97 (93 - 100)	40 (20%)
Moxifloxacin	HIV test result: **Negative**	427	22	92 (84 - 99)	513	15	97 (93 - 100)	91 (9%)
Streptomycin	Semi-quantitative Xpert result: **medium or high**	446	10	97 (95 - 100)	195	25	88 (81 - 96)	75 (10%)
Streptomycin	Semi-quantitative Xpert result: **low or very low**	36	1	96 (88 - 100)	25	5	77 (59 - 95)	64 (49%)
Streptomycin	HIV test result: **Positive**	23	0		47	2	94 (84 - 100)	26 (27%)
Streptomycin	HIV test result: **Negative**	313	3		106	20	83 (72 - 95)	92 (17%)
Amikacin	Semi-quantitative Xpert result: **medium or high**	97	10	*87 (74 - 100)	1183	8	*99 (98 - 100)	98 (7%)
Amikacin	Semi-quantitative Xpert result: **low or very low**	9	1	*91 (72 - 100)	157	2	*98 (96 - 100)	99 (37%)
Amikacin	HIV test result: **Positive**	4	0		148	0		44 (22%)
Amikacin	HIV test result: **Negative**	78	10		848	8		124 (12%)
Bedaquiline	Semi-quantitative Xpert result: **medium or high**	23	5	66 (37 - 95)	619	12	*97 (94 - 100)	31 (4%)
Bedaquiline	Semi-quantitative Xpert result: **low or very low**	3	0		64	2	*96 (91 - 100)	61 (47%)
Bedaquiline	HIV test result: **Positive**	8	0		67	4	*93 (86 - 100)	19 (19%)
Bedaquiline	HIV test result: **Negative**	18	4		441	9	*97 (96 - 99)	61 (11%)
Clofazimine	Semi-quantitative Xpert result: **medium or high**	27	7		699	15	97 (95 - 100)	35 (4%)
Clofazimine	Semi-quantitative Xpert result: **low or very low**	2	0		72	3	96 (91 - 100)	61 (44%)
Clofazimine	HIV test result: **Positive**	9	0		67	3	96 (92 - 100)	19 (19%)
Clofazimine	HIV test result: **Negative**	18	3		442	9	98 (97 - 99)	61 (11%)
Linezolid	Semi-quantitative Xpert result: **medium or high**	27	24	*73 (42 - 100)	1275	2	100 (100 - 100)	55 (4%)
Linezolid	Semi-quantitative Xpert result: **low or very low**	2	8	*35 (0 - 85)	136	0		114 (44%)
Linezolid	HIV test result: **Positive**	0	0		151	0		45 (23%)
Linezolid	HIV test result: **Negative**	21	30		908	2	100 (99 - 100)	105 (10%)

* Model not controlled for rifampicin resistance by MTB/RIF Xpert

** Model not controlled for MTB/RIF Xpert semi-quantitative result

***No data on HIV amoing the data for Ethambutol as only from TBSeq

Detailed results addressing each PICO in turn are reported below. For each PICO, the number of TP, FP, TN and FN samples are shown that were included in the model, separated by study. Accompanying forest plots are depicted. For each PICO there are also bar charts showing the proportion of studies, and proportion of samples across all studies, that were assessed to be at 'low risk', 'unclear risk', or 'high risk' of bias for any of the QUADAS2 domains. Similarly, a second histogram is included showing the proportion of studies assessed to be 'low risk', 'unclear risk', or 'high risk' for any applicability concerns within the QUADAS2 framework. All numbers are presented inclusive of any duplicate samples. Finally, the final GRADE tables are also included.

PICO 1 - tNGS as initial test for drug resistance in patients with bacteriologically-confirmed TB for RIF, INH, LFX, MFX, PZA, EMB:

PICO 1.1: Should tNGS as the initial test be used to diagnose drug resistance to rifampicin (RIF) (composite) in patients with bacteriologically confirmed pulmonary TB disease?

Nine studies with 1436 samples were included in the model for sensitivity, and 7 studies with 271 samples for specificity. Prevalence of resistance to rifampicin (composite) among samples included in the model was 84% (95% CI 82-86%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 12.0% (95% CI 10.5-13.6).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 93% (95% confidence interval (CI) 87%-99%)

Pooled specificity was 96% (95% confidence interval (CI) 89%-100%)

Negative predictive value was 85% (95% confidence interval (CI) 78%-91%)

Positive predictive value was 99% (95% confidence interval (CI) 97%-100%)

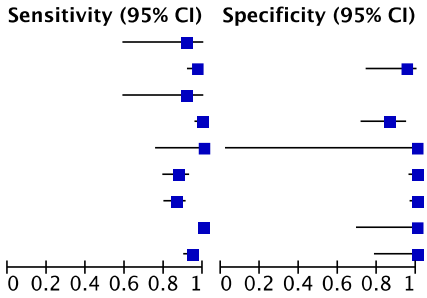
GRADE assessment:
115 observations were dropped by the model (already excluded in the numbers below) because for NanoTB the variable 'duplicate=2' predicted the outcome perfectly for 115 TN results. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. We further downgraded for specificity due to imprecision (the 95% confidence intervals around the pooled estimate were >10 percentage points). Quality of evidence was thereby assessed to be moderate for sensitivity and low for specificity.

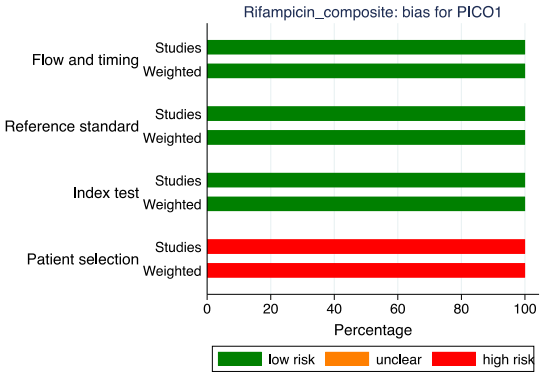
Data by study:

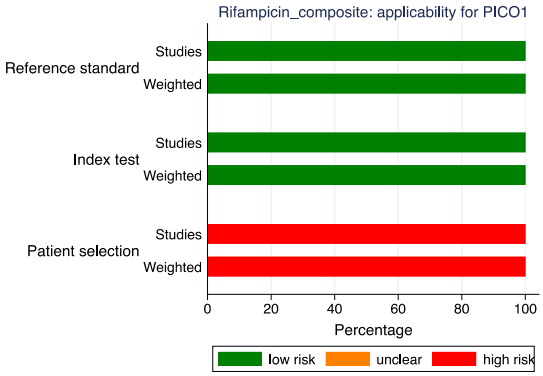
Rifampicin_composite

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama - Benin	10	0	1	0	0.91 [0.59, 1.00]	Not estimable
Diama - Cameroon	114	1	4	18	0.97 [0.92, 0.99]	0.95 [0.74, 1.00]
Diama - Ethiopia	10	0	1	0	0.91 [0.59, 1.00]	Not estimable
Diama - Guinea	116	6	1	36	0.99 [0.95, 1.00]	0.86 [0.71, 0.95]
Diama - Mali	13	0	0	1	1.00 [0.75, 1.00]	1.00 [0.03, 1.00]
Diama - Rwanda	93	0	14	85	0.87 [0.79, 0.93]	1.00 [0.96, 1.00]
FIND - Georgia	139	0	23	99	0.86 [0.79, 0.91]	1.00 [0.96, 1.00]
FIND - India	708	0	3	10	1.00 [0.99, 1.00]	1.00 [0.69, 1.00]
FIND - South Africa	175	0	11	15	0.94 [0.90, 0.97]	1.00 [0.78, 1.00]



QUADAS-2 assessment:





GRADE assessment:

Question: Should TNGS as the initial test be used to diagnose drug resistance to rifampin (RIF) (composite) in patients with bacteriologically confirmed pulmonary TB disease?

Sensitivity	0.93 (95% CI: 0.87 to 0.99)
Specificity	0.96 (95% CI: 0.89 to 1.00)

Prevalences	2%	10%	15%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of2%	pre-test probability of10%	pre-test probability of15%	
True positives (patients with drug resistance to rifampin (RIF) (composite))	9 studies 1436 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	19 (17 to 20)	93 (87 to 99)	140 (131 to 149)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to rifampin (RIF) (composite))								1 (0 to 3)	7 (1 to 13)	10 (1 to 19)	
True negatives (patients without drug resistance to rifampin (RIF) (composite))	7 studies 271 patients[b]	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	serious[c]	none	941 (872 to 980)	864 (801 to 900)	816 (757 to 850)	⨁⨁◯◯ Low
False positives (patients incorrectly classified as having drug resistance to rifampin (RIF) (composite))								39 (0 to 108)	36 (0 to 99)	34 (0 to 93)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to rifampicin (composite) across data used in the model was 83% (CI 81% to 85%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.
b. 115 observations from ONT dropped by model as variable 'duplicate=2' (i.e. ONT) predicts the outcome perfectly (115 TN results)
c. 95% confidence interval for specificity spans >10%

PICO 1.2: Should tNGS as the initial test be used to diagnose drug resistance to rifampin (RIF) (pDST) in patients with bacteriologically confirmed pulmonary TB disease?

Thirteen studies with 961 samples were included in the model for sensitivity, and 12 studies with 425 samples for specificity. Prevalence of resistance to rifampicin (pDST) among samples included in the model was 69% (95% CI 67-72%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 9.9% (95% CI 8.4-

11.6).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 99% (95% confidence interval (CI) 97%-100%)

Pooled specificity was 81% (95% confidence interval (CI) 69%-93%)

Negative predictive value was 98% (95% confidence interval (CI) 96%-99%)

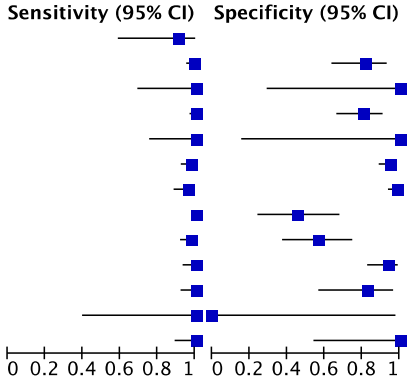
Positive predictive value was 94% (95% confidence interval (CI) 92%-96%)

GRADE assessment:
We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. We further downgraded for specificity for inconsistency (some of the data from FIND were outlying – see forest plot below), and for imprecision (the 95% confidence intervals around the pooled estimate were 24 percentage points wide). Quality of evidence was thereby assessed to be moderate for sensitivity and very low for specificity.

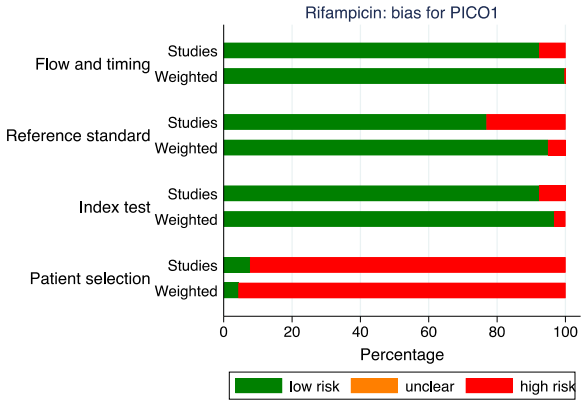
Data by study:

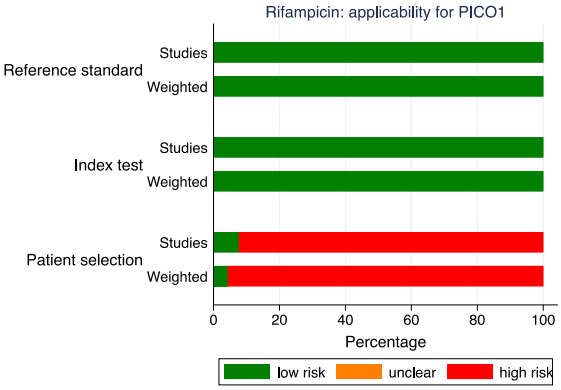
Rifampicin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama – Benin	10	0	1	0	0.91 [0.59, 1.00]	Not estimable
Diama – Cameroon	108	6	1	26	0.99 [0.95, 1.00]	0.81 [0.64, 0.93]
Diama – Ethiopia	10	0	0	3	1.00 [0.69, 1.00]	1.00 [0.29, 1.00]
Diama – Guinea	110	9	0	37	1.00 [0.97, 1.00]	0.80 [0.66, 0.91]
Diama – Mali	13	0	0	2	1.00 [0.75, 1.00]	1.00 [0.16, 1.00]
Diama – Rwanda	87	6	2	105	0.98 [0.92, 1.00]	0.95 [0.89, 0.98]
FIND – Georgia	69	2	3	106	0.96 [0.88, 0.99]	0.98 [0.93, 1.00]
FIND – India	333	12	0	10	1.00 [0.99, 1.00]	0.45 [0.24, 0.68]
FIND – South Africa	82	13	2	17	0.98 [0.92, 1.00]	0.57 [0.37, 0.75]
icddr,b – Bangladesh	51	3	0	44	1.00 [0.93, 1.00]	0.94 [0.82, 0.99]
IJTLD2022	44	3	0	14	1.00 [0.92, 1.00]	0.82 [0.57, 0.96]
OSR – Eritrea	4	1	0	0	1.00 [0.40, 1.00]	0.00 [0.00, 0.97]
Tuberculosis2021	31	0	0	6	1.00 [0.89, 1.00]	1.00 [0.54, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.99 (95% CI: 0.97 to 1.00)
Specificity	0.81 (95% CI: 0.69 to 0.93)

Prevalences	2%	10%	15%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of2%	pre-test probability of10%	pre-test probability of15%	
True positives (patients with drug resistance to rifampin (RIF) (pDST))	13 studies 961 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious[a]	none	20 (19 to 20)	99 (97 to 100)	149 (146 to 150)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to rifampin (RIF) (pDST))								0 (0 to 1)	1 (0 to 3)	1 (0 to 4)	
True negatives (patients without drug resistance to rifampin (RIF) (pDST))	12 studies 425 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	serious[b]	serious[c]	none	794 (676 to 911)	729 (621 to 837)	689 (586 to 791)	⨁◯◯◯ Very low
False positives (patients incorrectly classified as having drug resistance to rifampin (RIF) (pDST))								186 (69 to 304)	171 (63 to 279)	161 (59 to 264)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to rifampicin (pDST) across data used in the model was 69% (CI 67% to 72%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.
b. Two of the larger studies perform poorly for specificity
c. 95% confidence interval for specificity spans 24%

PICO 1.3: Should tNGS as the initial test be used to diagnose drug resistance to isoniazid (INH) (pDST) in patients with bacteriologically confirmed pulmonary TB disease?

Twelve studies with 1440 samples were included in the model for sensitivity, and 12 studies with 517 samples for specificity. Prevalence of resistance to isoniazid among samples included in the model was 74% (95% CI 72-76%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 14.6% (95% CI 13.0-16.2).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 96% (95% confidence interval (CI) 93%-99%)

Pooled specificity was 97% (95% confidence interval (CI) 95%-99%)

Negative predictive value was 91% (95% confidence interval (CI) 88%-94%)

Positive predictive value was 99% (95% confidence interval (CI) 98%-100%)

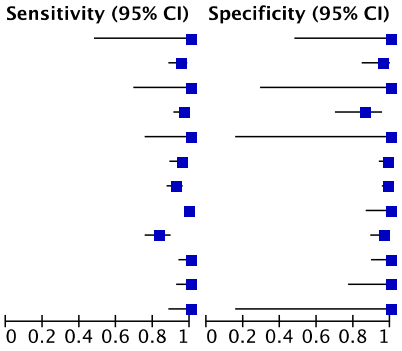
GRADE assessment

We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. Quality of evidence was thereby assessed to be moderate for both sensitivity and specificity.

Data by study:

Isoniazid

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama – Benin	5	0	0	5	1.00 [0.48, 1.00]	1.00 [0.48, 1.00]
Diama – Cameroon	89	2	5	41	0.95 [0.88, 0.98]	0.95 [0.84, 0.99]
Diama – Ethiopia	10	0	0	3	1.00 [0.69, 1.00]	1.00 [0.29, 1.00]
Diama – Guinea	104	5	4	30	0.96 [0.91, 0.99]	0.86 [0.70, 0.95]
Diama – Mali	13	0	0	2	1.00 [0.75, 1.00]	1.00 [0.16, 1.00]
Diama – Rwanda	82	2	4	105	0.95 [0.89, 0.99]	0.98 [0.93, 1.00]
FIND – Georgia	154	3	13	171	0.92 [0.87, 0.96]	0.98 [0.95, 1.00]
FIND – India	697	0	6	25	0.99 [0.98, 1.00]	1.00 [0.86, 1.00]
FIND – South Africa	103	3	21	72	0.83 [0.75, 0.89]	0.96 [0.89, 0.99]
icddr,b – Bangladesh	55	0	0	32	1.00 [0.94, 1.00]	1.00 [0.89, 1.00]
IJTLD2022	46	0	0	14	1.00 [0.92, 1.00]	1.00 [0.77, 1.00]
Tuberculosis2021	29	0	0	2	1.00 [0.88, 1.00]	1.00 [0.16, 1.00]



QUADAS-2 assessment:

GRADE table:

Sensitivity	0.96 (95% CI: 0.93 to 0.99)
Specificity	0.97 (95% CI: 0.95 to 0.99)

Prevalences	2%	10%	15%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of2%	pre-test probability of10%	pre-test probability of15%	
True positives (patients with drug resistance to isoniazid (INH) (pDST))	12 studies 1440 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	19 (19 to 20)	96 (93 to 99)	144 (140 to 149)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to isoniazid (INH) (pDST))								1 (0 to 1)	4 (1 to 7)	6 (1 to 10)	
True negatives (patients without drug resistance to isoniazid (INH) (pDST))	12 studies 517 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	951 (931 to 970)	873 (855 to 891)	825 (808 to 842)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having drug resistance to isoniazid (INH) (pDST))								29 (10 to 49)	27 (9 to 45)	25 (8 to 42)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to isoniazid across data used in the model was 74% (CI 72% to 76%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.

PICO 1.4: Should tNGS as the initial test be used to diagnose drug resistance to levofloxacin (LFX) (pDST) in patients with bacteriologically confirmed pulmonary TB disease?

Six studies with 654 samples were included in the model for sensitivity, and 7 studies with 913 samples for specificity. Prevalence of resistance to levofloxacin among samples included in the model was 42% (95% CI 39-44%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 9.2% (95% CI 7.8-10.7).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 94% (95% confidence interval (CI) 88%-100%)

Pooled specificity was 96% (95% confidence interval (CI) 93%-99%)

Negative predictive value was 97% (95% confidence interval (CI) 96%-98%)

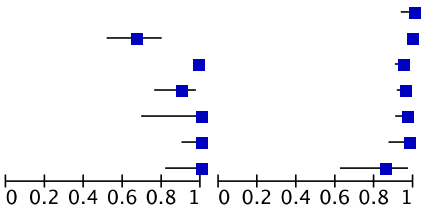
Positive predictive value was 93% (95% confidence interval (CI) 88%-98%)

GRADE assessment:
We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. We further downgraded for sensitivity for inconsistency (sone of the FIND studies was outlying – see forest plot below). Quality of evidence was thereby assessed to be low for sensitivity and moderate for specificity.

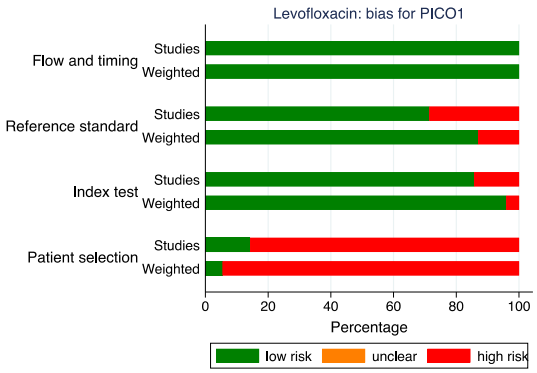
Data by study:

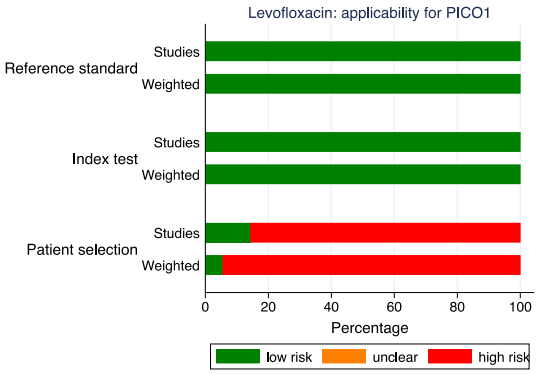
Levofloxacin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama – Cameroon	0	0	0	52	Not estimable	1.00 [0.93, 1.00]
FIND – Georgia	32	3	16	316	0.67 [0.52, 0.80]	0.99 [0.97, 1.00]
FIND – India	496	13	9	209	0.98 [0.97, 0.99]	0.94 [0.90, 0.97]
FIND – South Africa	35	8	4	165	0.90 [0.76, 0.97]	0.95 [0.91, 0.98]
icddr,b – Bangladesh	10	3	0	84	1.00 [0.69, 1.00]	0.97 [0.90, 0.99]
IJTLD2022	34	1	0	39	1.00 [0.90, 1.00]	0.97 [0.87, 1.00]
Tuberculosis2021	18	3	0	17	1.00 [0.81, 1.00]	0.85 [0.62, 0.97]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.94 (95% CI: 0.88 to 1.00)
Specificity	0.96 (95% CI: 0.93 to 0.99)

Prevalences	1%	5%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of5%	pre-test probability of10%	
True positives (patients with drug resistance to levofloxacin (LFX) (pDST))	6 studies 654 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	serious[b]	not serious	none	9 (9 to 10)	47 (44 to 50)	94 (88 to 100)	⨁⨁◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to levofloxacin (LFX) (pDST))								1 (0 to 1)	3 (0 to 6)	6 (0 to 12)	
True negatives (patients without drug resistance to levofloxacin (LFX) (pDST))	7 studies 913 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	950 (921 to 980)	912 (884 to 941)	864 (837 to 891)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having drug resistance to levofloxacin (LFX) (pDST))								40 (10 to 69)	38 (9 to 66)	36 (9 to 63)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to Levofloxacin across data used in the model was 42% (CI 39% to 44%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.
b. One of the larger studies performed much worse for sensitivity

PICO 1.5: Should tNGS as the initial test be used to diagnose drug resistance to moxifloxacin (MFX) (pDST) in patients with bacteriologically confirmed pulmonary TB disease?

Six studies with 652 samples were included in the model for sensitivity, and 8 studies with 921 samples for specificity. Prevalence of resistance to moxifloxacin among samples included in the model was 41% (95% CI 39-44%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 9.3% (95% CI 7.9-10.9).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 96% (95% confidence interval (CI) 92%-99%)

Pooled specificity was 96% (95% confidence interval (CI) 93%-100%)

Negative predictive value was 97% (95% confidence interval (CI) 96%-99%)

Positive predictive value was 94% (95% confidence interval (CI) 91%-98%)

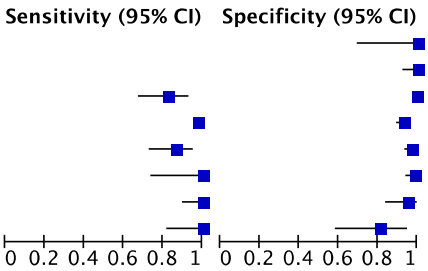
GRADE assessment:
We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias

as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. Quality of evidence was thereby assessed to be moderate for sensitivity and for specificity.

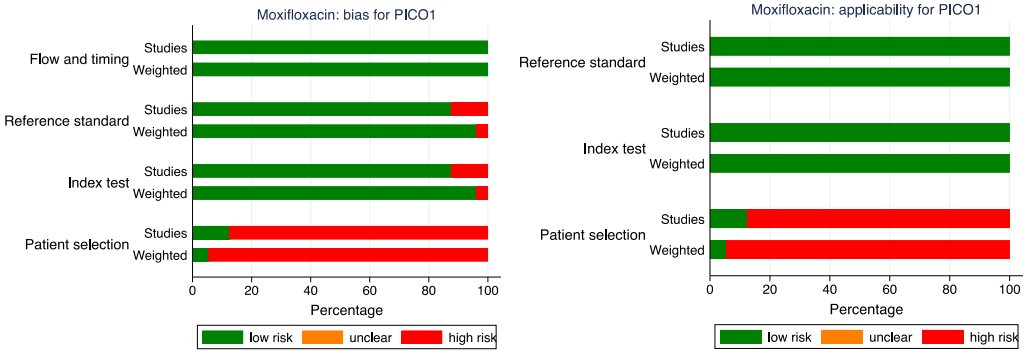
Data by study:

Moxifloxacin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama - Benin	0	0	0	10	Not estimable	1.00 [0.69, 1.00]
Diama - Cameroon	0	0	0	45	Not estimable	1.00 [0.92, 1.00]
FIND - Georgia	33	2	7	327	0.82 [0.67, 0.93]	0.99 [0.98, 1.00]
FIND - India	494	15	11	207	0.98 [0.96, 0.99]	0.93 [0.89, 0.96]
FIND - South Africa	38	5	6	163	0.86 [0.73, 0.95]	0.97 [0.93, 0.99]
icddr,b - Bangladesh	12	1	0	84	1.00 [0.74, 1.00]	0.99 [0.94, 1.00]
IJTLD2022	33	2	0	39	1.00 [0.89, 1.00]	0.95 [0.83, 0.99]
Tuberculosis2021	18	4	0	17	1.00 [0.81, 1.00]	0.81 [0.58, 0.95]



QUADAS-2 assessment:



GRADE table:

Sensitivity	0.96 (95% CI: 0.92 to 0.99)
Specificity	0.96 (95% CI: 0.93 to 1.00)

Prevalences	1%	5%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of5%	pre-test probability of10%	
True positives (patients with drug resistance to moxifloxacin (MFX) (pDST))	6 studies 652 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	10 (9 to 10)	48 (46 to 50)	96 (92 to 99)	⊕⊕⊕◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to moxifloxacin (MFX) (pDST))								0 (0 to 1)	2 (0 to 4)	4 (1 to 8)	
True negatives (patients without drug resistance to moxifloxacin (MFX) (pDST))	8 studies 921 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	950 (921 to 990)	912 (884 to 950)	864 (837 to 900)	⊕⊕⊕◯ Moderate
False positives (patients incorrectly classified as having drug resistance to moxifloxacin (MFX) (pDST))								40 (0 to 69)	38 (0 to 66)	36 (0 to 63)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to Moxifloxacin across data used in the model was 41% (CI 39% to 44%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.

PICO 1.6: Should tNGS as the initial test be used to diagnose drug resistance to pyrazinamide (PZA) (composite) in patients with bacteriologically confirmed pulmonary TB disease?

Three studies with 346 samples were included in the model for sensitivity, and 3 studies with 269 samples for specificity. Prevalence of resistance to pyrazinamide (composite) among samples included in the model was 56% (95% CI 52-60%). The proportion of

samples with no tNGS result reported (the 'indeterminate rate') was 17.6% (95% CI 14.6-20.8).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 88% (95% confidence interval (CI) 85%-92%)

Pooled specificity was 99% (95% confidence interval (CI) 97%-100%)

Negative predictive value was 88% (95% confidence interval (CI) 81%-95%)

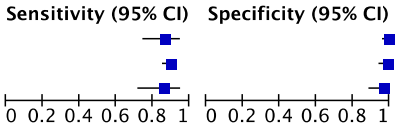
Positive predictive value was 99% (95% confidence interval (CI) 98%-100%)

GRADE assessment:
We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. Quality of evidence was thereby assessed to be moderate for sensitivity and for specificity.

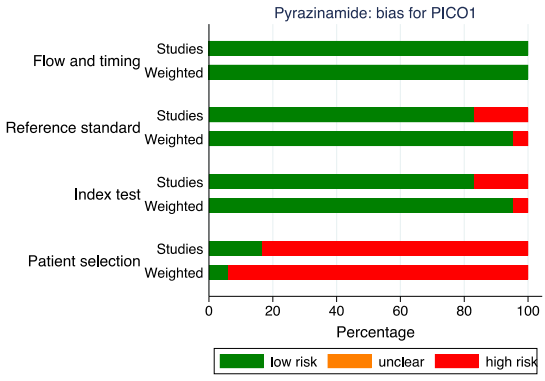
Data by study:

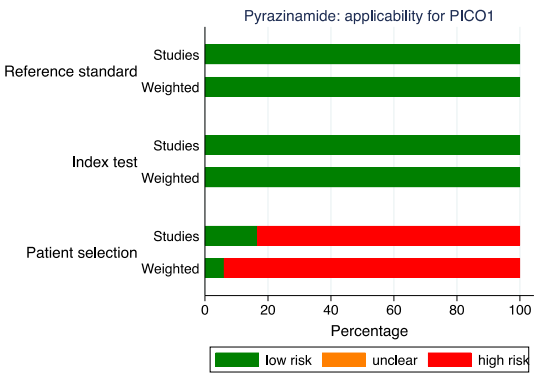
Pyrazinamide_composite

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	45	1	7	123	0.87 [0.74, 0.94]	0.99 [0.96, 1.00]
FIND – India	225	1	27	85	0.89 [0.85, 0.93]	0.99 [0.94, 1.00]
FIND – South Africa	36	2	6	57	0.86 [0.71, 0.95]	0.97 [0.88, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.88 (95% CI: 0.85 to 0.92)
Specificity	0.99 (95% CI: 0.97 to 1.00)

Prevalences	1%	3%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of10%	
True positives (patients with drug resistance to pyrazinamide (PZA) (composite))	3 studies 346 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	9 (9 to 9)	26 (26 to 28)	88 (85 to 92)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to pyrazinamide (PZA) (composite))								1 (1 to 1)	4 (2 to 4)	12 (8 to 15)	
True negatives (patients without drug resistance to pyrazinamide (PZA) (composite))	3 studies 269 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	980 (960 to 990)	960 (941 to 970)	891 (873 to 900)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having drug resistance to pyrazinamide (PZA) (composite))								10 (0 to 30)	10 (0 to 29)	9 (0 to 27)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to Pyrazinamide (composite) across data used in the model was 56% (CI 52% to 60%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.

PICO 1.7: Should tNGS as the initial test be used to diagnose drug resistance to pyrazinamide (PZA) (pDST) in patients with bacteriologically confirmed pulmonary TB disease?

Six studies with 425 samples were included in the model for sensitivity, and 6 studies with 379 samples for specificity. Prevalence of resistance to pyrazinamide (pDST) among samples included in the model was 52% (95% CI 49-56%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 14.7% (95% CI 12.3-17.3).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 85% (95% confidence interval (CI) 80%-90%)

Pooled specificity was 94% (95% confidence interval (CI) 92%-96%)

Negative predictive value was 84% (95% confidence interval (CI) 75%-93%)

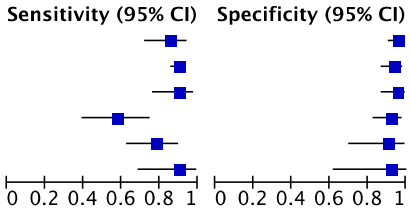
Positive predictive value was 92% (95% confidence interval (CI) 88%-97%)

GRADE assessment:
We downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. We further downgraded for sensitivity for inconsistency (the study from Bangladesh, icddr,b, was outlying – see forest plot below). Quality of evidence was thereby assessed to be low for sensitivity and moderate for specificity.

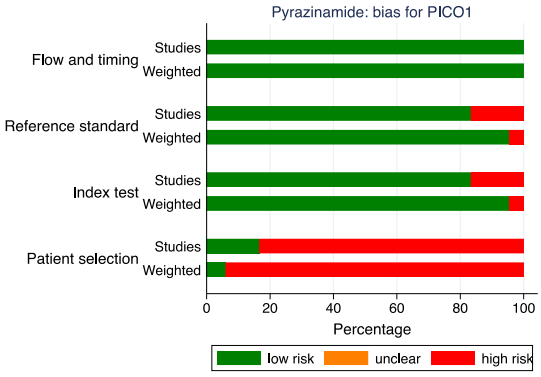
Data by study:

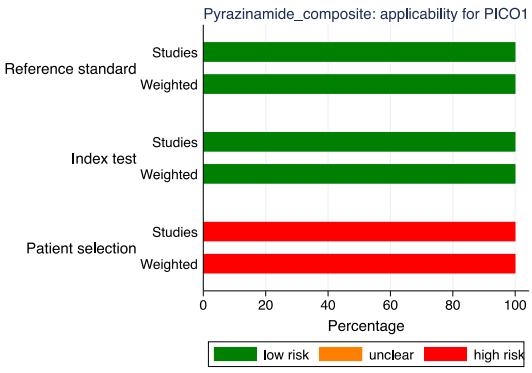
Pyrazinamide

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND - Georgia	40	6	7	123	0.85 [0.72, 0.94]	0.95 [0.90, 0.98]
FIND - India	220	6	25	87	0.90 [0.85, 0.93]	0.94 [0.86, 0.98]
FIND - South Africa	35	3	4	59	0.90 [0.76, 0.97]	0.95 [0.87, 0.99]
icddr,b - Bangladesh	19	5	14	57	0.58 [0.39, 0.75]	0.92 [0.82, 0.97]
IJTLD2022	32	2	9	19	0.78 [0.62, 0.89]	0.90 [0.70, 0.99]
Tuberculosis2021	18	1	2	11	0.90 [0.68, 0.99]	0.92 [0.62, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.85 (95% CI: 0.80 to 0.90)
Specificity	0.94 (95% CI: 0.92 to 0.96)

Prevalences	1%	3%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of10%	
True positives (patients with drug resistance to pyrazinamide (PZA) (pDST))	6 studies 425 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	serious[b]	not serious	none	9 (8 to 9)	26 (24 to 27)	85 (80 to 90)	⨁⨁◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to pyrazinamide (PZA) (pDST))								1 (1 to 2)	4 (3 to 6)	15 (10 to 20)	
True negatives (patients without drug resistance to pyrazinamide (PZA) (pDST))	6 studies 379 patients	cross-sectional (cohort type accuracy study)	not serious[a]	serious[a]	not serious	not serious	none	931 (911 to 950)	912 (892 to 931)	846 (828 to 864)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having drug resistance to pyrazinamide (PZA) (pDST))								59 (40 to 79)	58 (39 to 78)	54 (36 to 72)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to pyrazinamide (pDST) across data used in the model was 52% (CI 49% to 56%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.
b. One study had a much lower sensitivity

PICO 1.8: Should tNGS as the initial test be used to diagnose drug resistance to ethambutol (EMB) (composite) in patients with bacteriologically confirmed pulmonary TB disease?

Four studies with 432 samples were included in the model for sensitivity, and 4 studies with 268 samples for specificity. Prevalence of resistance to ethambutol (composite) among samples included in the model was 62% (95% CI 58-65%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 16.3% (95% CI 13.5-19.2).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 96% (95% confidence interval (CI) 94%-98%)

Pooled specificity was 99% (95% confidence interval (CI) 98%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear.

Negative predictive value was 93% (95% confidence interval (CI) 89%-97%)

Positive predictive value was 100% (95% confidence interval (CI) 99%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear.

GRADE assessment:

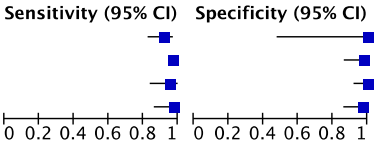
The model for specificity did not control for rifampicin resistance as this variable was collinear in the original model. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for risk of bias for both sensitivity and specificity as difference samples were used for the index (tNGS) and reference tests. We also downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. Quality of evidence was thereby assessed to be low for both sensitivity and specificity.

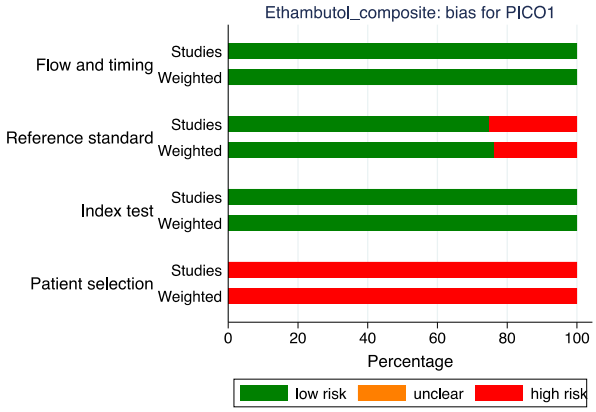
Data by study:

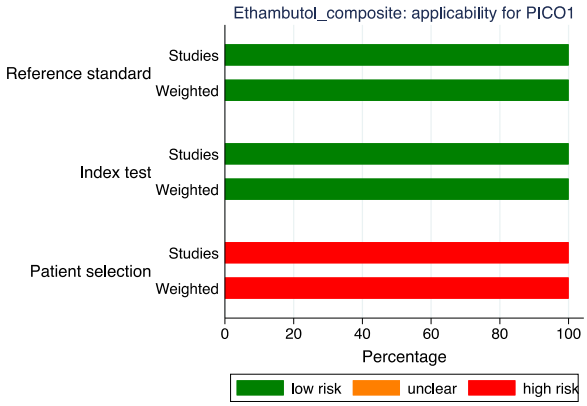
Ethambutol_composite

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	64	0	6	5	0.91 [0.82, 0.97]	1.00 [0.48, 1.00]
FIND – India	275	1	9	37	0.97 [0.94, 0.99]	0.97 [0.86, 1.00]
FIND – South Africa	39	0	2	43	0.95 [0.83, 0.99]	1.00 [0.92, 1.00]
TBSeq – China	36	1	1	36	0.97 [0.86, 1.00]	0.97 [0.86, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.96 (95% CI: 0.94 to 0.98)
Specificity	0.99 (95% CI: 0.98 to 1.00)

Prevalences	1%	3%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of10%	
True positives (patients with drug resistance to ethambutol (EMB) (composite))	4 studies 432 patients	cross-sectional (cohort type accuracy study)	serious[a,b]	serious[a]	not serious	not serious	none	10 (9 to 10)	29 (28 to 29)	96 (94 to 98)	⨁⨁◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to ethambutol (EMB) (composite))								0 (0 to 1)	1 (1 to 2)	4 (2 to 6)	
True negatives (patients without drug resistance to ethambutol (EMB) (composite))	4 studies 268 patients[c]	cross-sectional (cohort type accuracy study)	serious[a,b]	serious[a]	not serious	not serious	none	980 (970 to 990)	960 (951 to 970)	891 (882 to 900)	⨁⨁◯◯ Low
False positives (patients incorrectly classified as having drug resistance to ethambutol (EMB) (composite))								10 (0 to 20)	10 (0 to 19)	9 (0 to 18)	

Explanations

a. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to ethambutol (composite) across data used in the model was 62% (CI 58% to 65%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.
b. Different samples used for tNGS and reference test
c. The model does not control for rifampicin resistance as this variable was collinear in the original model.

PICO 1.9: Should tNGS as the initial test be used to diagnose drug resistance to ethambutol (EMB) (pDST) in patients with bacteriologically confirmed pulmonary TB disease?

One study with 100 samples was included in the model for sensitivity, and one study with 334 samples for specificity. Prevalence of resistance to ethambutol (pDST) among samples included in the model was 23% (95% CI 19-27%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 0% (95% CI 0-1).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 88% (95% confidence interval (CI) 82%-94%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as variable was collinear.

Pooled specificity was 94% (95% confidence interval (CI) 91%-97%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear.

Negative predictive value was 96% (95% confidence interval (CI) 93%-99%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear.

Positive predictive value was 83% (95% confidence interval (CI) 75%-90%)

Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear.

GRADE assessment:

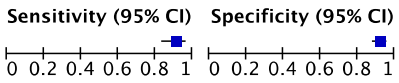
The models for sensitivity and specificity do not control for rifampicin resistance as this variable was collinear in the original model. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for risk of bias for both sensitivity and specificity as difference samples were used for the index (tNGS) and reference tests. We also downgraded one for indirectness for both sensitivity and specificity as all studies were enriched for rifampicin resistance, leading to applicability concerns, but not for risk of bias as sensitivity and specificity should not be unduly impacted by a change in the prevalence of resistance. We further downgraded for indirectness as all the data were from just a single study and generalizability could be limited. Quality of evidence was thereby assessed to be very low for both sensitivity and specificity.

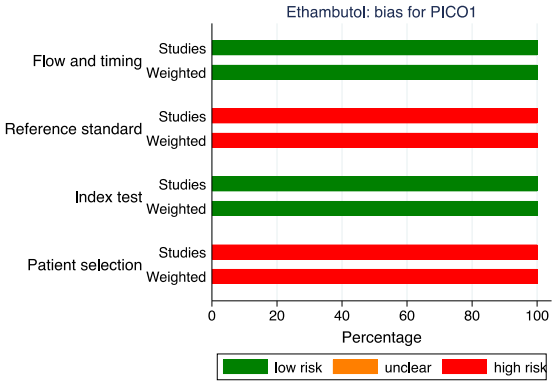
Data by study:

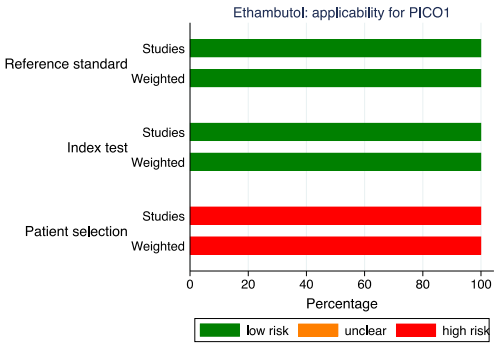
Ethambutol

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
TBSeq - China	81	17	8	196	0.91 [0.83, 0.96]	0.92 [0.88, 0.95]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.88 (95% CI: 0.82 to 0.94)
Specificity	0.94 (95% CI: 0.91 to 0.97)

Prevalences	1%	3%	10%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of10%	
True positives (patients with drug resistance to ethambutol (EMB) (pDST))	1 studies 100 patients[a]	cross-sectional (cohort type accuracy study)	serious[b,c]	very serious[b,d]	not serious	not serious	none	9 (8 to 9)	26 (25 to 28)	88 (82 to 94)	⊕◯◯◯ Very low
False negatives (patients incorrectly classified as not having drug resistance to ethambutol (EMB) (pDST))								1 (1 to 2)	4 (2 to 5)	12 (6 to 18)	
True negatives (patients without drug resistance to ethambutol (EMB) (pDST))	1 studies 334 patients[a]	cross-sectional (cohort type accuracy study)	serious[b,c]	very serious[b,d]	not serious	not serious	none	931 (901 to 960)	912 (883 to 941)	846 (819 to 873)	⊕◯◯◯ Very low
False positives (patients incorrectly classified as having drug resistance to ethambutol (EMB) (pDST))								59 (30 to 89)	58 (29 to 87)	54 (27 to 81)	

Explanations

a. The model does not control for rifampicin resistance as this variable was collinear in the original model.
b. All studies enriched for samples that were rifampicin resistant. Prevalence of resistance to ethambutol (pDST) across data used in the model was 23% (CI 19% to 27%). However, prevalence should not significantly impact sensitivity or specificity, therefore not downgraded for bias, just for indirectness.
c. Different samples were used for tNGS and reference assay
d. Only one study (from China). Downgraded by one as may not be generalisable.

PICO 2 - TNGS as test for drug resistance in patients with bac-confirmed rif-resistant TB (RR-TB) for INH, LFX, MFX, PZA, BDQ, LZD, CFZ, AMK, EMB, ETO, PTO, STR

PICO 2.1: Should tNGS be used to diagnose drug resistance to isoniazid (INH) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Twelve studies with 1440 samples were included in the model for sensitivity, and 12 studies with 517 samples for specificity. Prevalence of resistance to isoniazid among samples included in the model was 74% (95% CI 72-76%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 14.6% (95% CI 13.0-16.2).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 96% (95% confidence interval (CI) 94%-99%)

Pooled specificity was 96% (95% confidence interval (CI) 92%-100%)

Negative predictive value was 76% (95% confidence interval (CI) 67%-84%)

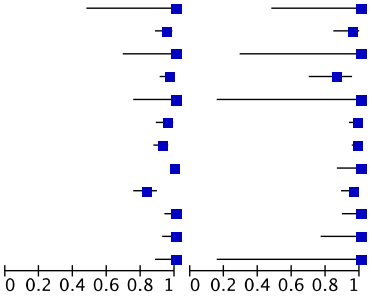
Positive predictive value was 100% (95% confidence interval (CI) 99%-100%)

GRADE assessment:
We did not downgrade for any categories. Quality of evidence was thereby assessed to be high for both sensitivity and specificity.

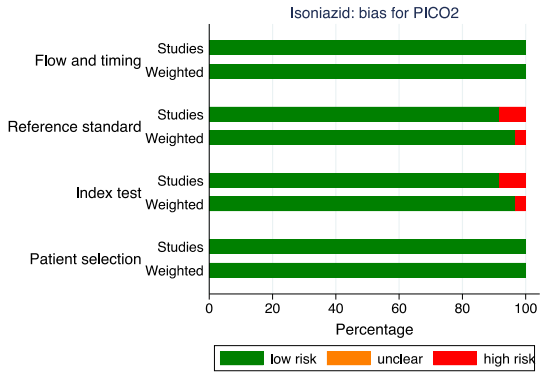
Data by study:

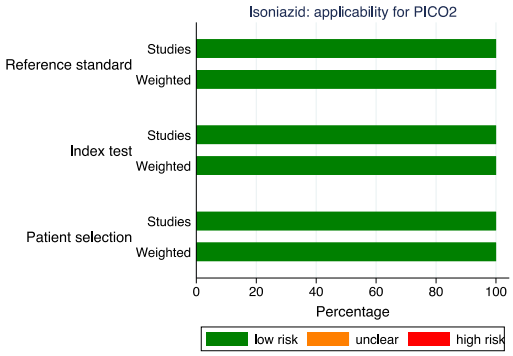
Isoniazid

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama – Benin	5	0	0	5	1.00 [0.48, 1.00]	1.00 [0.48, 1.00]
Diama – Cameroon	89	2	5	41	0.95 [0.88, 0.98]	0.95 [0.84, 0.99]
Diama – Ethiopia	10	0	0	3	1.00 [0.69, 1.00]	1.00 [0.29, 1.00]
Diama – Guinea	104	5	4	30	0.96 [0.91, 0.99]	0.86 [0.70, 0.95]
Diama – Mali	13	0	0	2	1.00 [0.75, 1.00]	1.00 [0.16, 1.00]
Diama – Rwanda	82	2	4	105	0.95 [0.89, 0.99]	0.98 [0.93, 1.00]
FIND – Georgia	154	3	13	171	0.92 [0.87, 0.96]	0.98 [0.95, 1.00]
FIND – India	697	0	6	25	0.99 [0.98, 1.00]	1.00 [0.86, 1.00]
FIND – South Africa	103	3	21	72	0.83 [0.75, 0.89]	0.96 [0.89, 0.99]
icddr,b – Bangladesh	55	0	0	32	1.00 [0.94, 1.00]	1.00 [0.89, 1.00]
IJTLD2022	46	0	0	14	1.00 [0.92, 1.00]	1.00 [0.77, 1.00]
Tuberculosis2021	29	0	0	2	1.00 [0.88, 1.00]	1.00 [0.16, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.96 (95% CI: 0.94 to 0.99)
Specificity	0.96 (95% CI: 0.92 to 1.00)

Prevalences	60%	75%	90%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of60%	pre-test probability of75%	pre-test probability of90%	
True positives (patients with drug resistance to isonazid (INH) (pDST))	12 studies 1440 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	576 (564 to 594)	720 (705 to 742)	864 (846 to 891)	⊕⊕⊕⊕ High
False negatives (patients incorrectly classified as not having drug resistance to isonazid (INH) (pDST))								24 (6 to 36)	30 (8 to 45)	36 (9 to 54)	
True negatives (patients without drug resistance to isonazid (INH) (pDST))	12 studies 517 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	384 (368 to 400)	240 (230 to 250)	96 (92 to 100)	⊕⊕⊕⊕ High
False positives (patients incorrectly classified as having drug resistance to isonazid (INH) (pDST))								16 (0 to 32)	10 (0 to 20)	4 (0 to 8)	

Explanations

a. Prevalence of resistance to isoniazid across data used in the model was 74% (CI 72% to 76%)

PICO 2.2: Should tNGS be used to diagnose drug resistance to levofloxacin (LFX) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Six studies with 654 samples were included in the model for sensitivity, and 7 studies with 913 samples for specificity. Prevalence of resistance to levofloxacin among samples included in the model was 42% (95% CI 39-44%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 9.2% (95% CI 7.8-10.7).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 96% (95% confidence interval (CI) 90%-100%)

Pooled specificity was 96% (95% confidence interval (CI) 93%-99%)

Negative predictive value was 98% (95% confidence interval (CI) 96%-99%)

Positive predictive value was 94% (95% confidence interval (CI) 90%-99%)

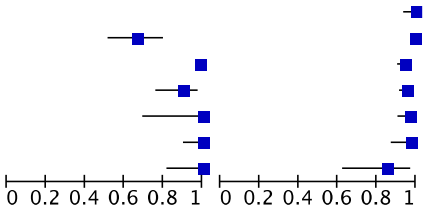
GRADE assessment:
For sensitivity we downgraded one for inconsistency as one of the FIND studies was an outlier (see forest plot below). Quality of evidence was thereby assessed to be moderate

for sensitivity and high for specificity.

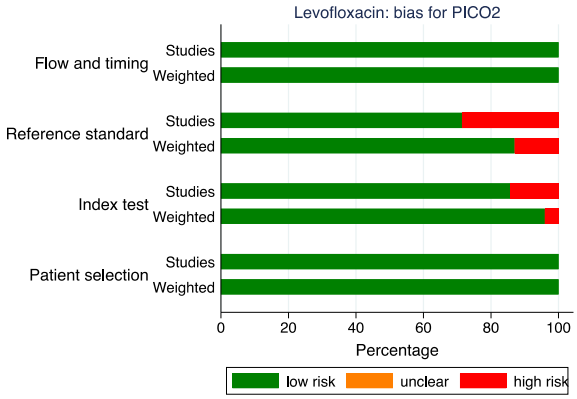
Data by study:

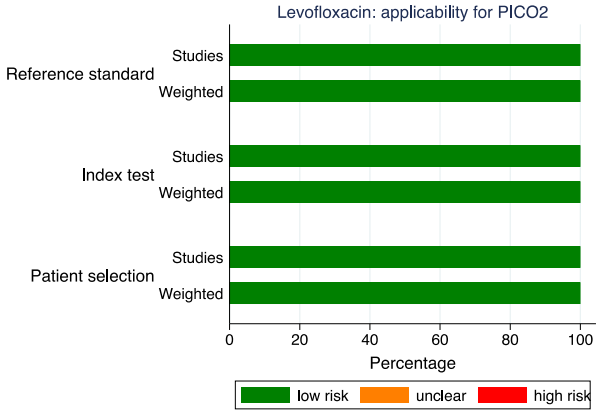
Levofloxacin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama – Cameroon	0	0	0	52	Not estimable	1.00 [0.93, 1.00]
FIND – Georgia	32	3	16	316	0.67 [0.52, 0.80]	0.99 [0.97, 1.00]
FIND – India	496	13	9	209	0.98 [0.97, 0.99]	0.94 [0.90, 0.97]
FIND – South Africa	35	8	4	165	0.90 [0.76, 0.97]	0.95 [0.91, 0.98]
icddr,b – Bangladesh	10	3	0	84	1.00 [0.69, 1.00]	0.97 [0.90, 0.99]
IJTLD2022	34	1	0	39	1.00 [0.90, 1.00]	0.97 [0.87, 1.00]
Tuberculosis2021	18	3	0	17	1.00 [0.81, 1.00]	0.85 [0.62, 0.97]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.96 (95% CI: 0.90 to 1.00)
Specificity	0.96 (95% CI: 0.93 to 0.99)

Prevalences	10%	30%	50%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of10%	pre-test probability of30%	pre-test probability of50%	
True positives (patients with drug resistance to levofloxacin (LFX) (pDST))	6 studies 654 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	serious[b]	not serious	none	96 (90 to 100)	288 (270 to 300)	480 (450 to 500)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to levofloxacin (LFX) (pDST))								4 (0 to 10)	12 (0 to 30)	20 (0 to 50)	
True negatives (patients without drug resistance to levofloxacin (LFX) (pDST))	7 studies 913 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	864 (837 to 891)	672 (651 to 693)	480 (465 to 495)	⨁⨁⨁⨁ High
False positives (patients incorrectly classified as having drug resistance to levofloxacin (LFX) (pDST))								36 (9 to 63)	28 (7 to 49)	20 (5 to 35)	

Explanations

a. Prevalence of resistance to levofloxacin across data used in the model was 42% (CI 39% to 44%)
b. One outlying study for sensitivity

PICO 2.3: Should tNGS be used to diagnose drug resistance to moxifloxacin (MFX) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Six studies with 652 samples were included in the model for sensitivity, and 8 studies with 921 samples for specificity. Prevalence of resistance to moxifloxacin among samples included in the model was 41% (95% CI 39-44%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 9.3% (95% CI 7.9-10.9).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 97% (95% confidence interval (CI) 94%-100%)

Pooled specificity was 95% (95% confidence interval (CI) 91%-99%)

Negative predictive value was 97% (95% confidence interval (CI) 96%-99%)

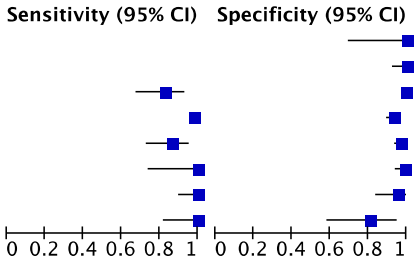
Positive predictive value was 95% (95% confidence interval (CI) 91%-99%)

GRADE assessment:
We did not downgrade for any categories. Quality of evidence was thereby assessed to be high for both sensitivity and specificity.

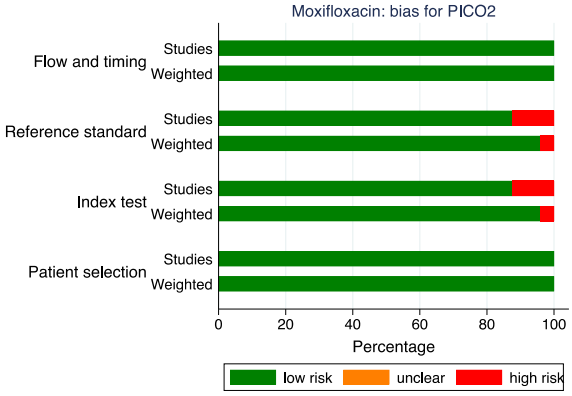
Data by study:

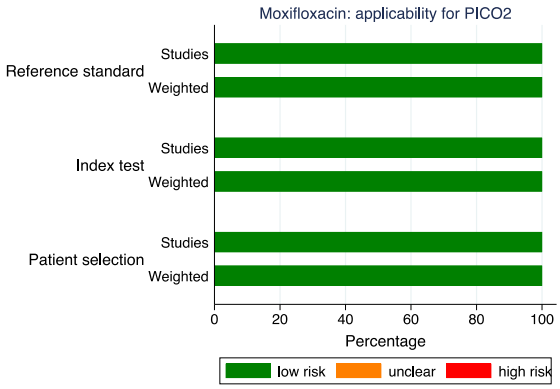
Moxifloxacin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama - Benin	0	0	0	10	Not estimable	1.00 [0.69, 1.00]
Diama - Cameroon	0	0	0	45	Not estimable	1.00 [0.92, 1.00]
FIND - Georgia	33	2	7	327	0.82 [0.67, 0.93]	0.99 [0.98, 1.00]
FIND - India	494	15	11	207	0.98 [0.96, 0.99]	0.93 [0.89, 0.96]
FIND - South Africa	38	5	6	163	0.86 [0.73, 0.95]	0.97 [0.93, 0.99]
icddr,b - Bangladesh	12	1	0	84	1.00 [0.74, 1.00]	0.99 [0.94, 1.00]
IJTLD2022	33	2	0	39	1.00 [0.89, 1.00]	0.95 [0.83, 0.99]
Tuberculosis2021	18	4	0	17	1.00 [0.81, 1.00]	0.81 [0.58, 0.95]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.97 (95% CI: 0.94 to 1.00)
Specificity	0.95 (95% CI: 0.91 to 0.99)

Prevalences	10%	30%	50%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of10%	pre-test probability of30%	pre-test probability of50%	
True positives (patients with drug resistance to moxifloxacin (MFX) (pDST))	6 studies 652 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	97 (94 to 100)	291 (282 to 300)	485 (470 to 500)	⊕⊕⊕⊕ High
False negatives (patients incorrectly classified as not having drug resistance to moxifloxacin (MFX) (pDST))								3 (0 to 6)	9 (0 to 18)	15 (0 to 30)	
True negatives (patients without drug resistance to moxifloxacin (MFX) (pDST))	8 studies 921 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	855 (819 to 891)	665 (637 to 693)	475 (455 to 495)	⊕⊕⊕⊕ High
False positives (patients incorrectly classified as having drug resistance to moxifloxacin (MFX) (pDST))								45 (9 to 81)	35 (7 to 63)	25 (5 to 45)	

Explanations

a. Prevalence of resistance to moxifloxcin across data used in the model was 41% (CI 39% to 44%)

PICO 2.4: Should tNGS be used to diagnose drug resistance to pyrazinamide (PZA) (composite) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Three studies with 346 samples were included in the model for sensitivity, and 3 studies with 269 samples for specificity. Prevalence of resistance to pyrazinamide (composite) among samples included in the model was 56% (95% CI 52-60%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 17.6% (95% CI 14.6-20.8).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 90% (95% confidence interval (CI) 87%-93%)

Pooled specificity was 99% (95% confidence interval (CI) 97%-100%)

Negative predictive value was 85% (95% confidence interval (CI) 75%-94%)

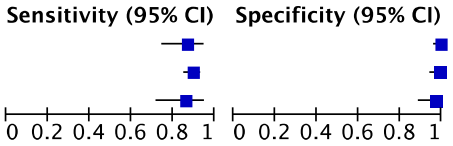
Positive predictive value was 99% (95% confidence interval (CI) 98%-100%)

GRADE assessment:
We did not downgrade for any categories. Quality of evidence was thereby assessed to be high for both sensitivity and specificity.

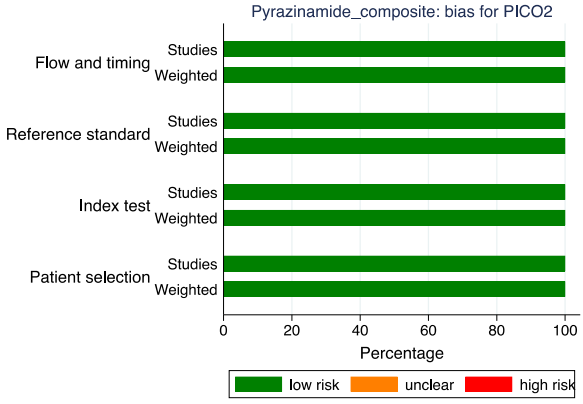
Data by study:

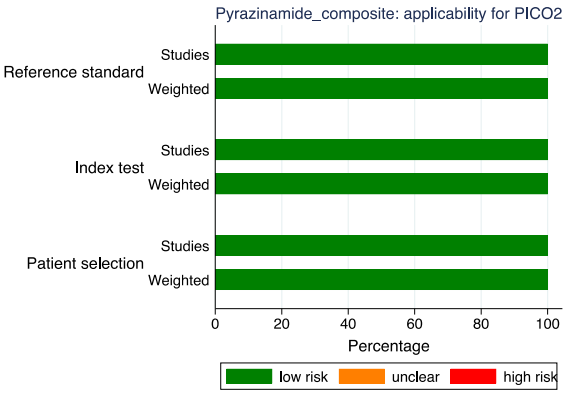
Pyrazinamide_composite

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	45	1	7	123	0.87 [0.74, 0.94]	0.99 [0.96, 1.00]
FIND – India	225	1	27	85	0.89 [0.85, 0.93]	0.99 [0.94, 1.00]
FIND – South Africa	36	2	6	57	0.86 [0.71, 0.95]	0.97 [0.88, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.90 (95% CI: 0.87 to 0.93)
Specificity	0.99 (95% CI: 0.97 to 1.00)

Prevalences	30%	50%	90%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of30%	pre-test probability of50%	pre-test probability of90%	
True positives (patients with drug resistance to pyrazinamide (PZA) (composite))	3 studies 346 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	270 (261 to 279)	450 (435 to 465)	810 (783 to 837)	⨁⨁⨁⨁ High
False negatives (patients incorrectly classified as not having drug resistance to pyrazinamide (PZA) (composite))								30 (21 to 39)	50 (35 to 65)	90 (63 to 117)	
True negatives (patients without drug resistance to pyrazinamide (PZA) (composite))	3 studies 269 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	693 (679 to 700)	495 (485 to 500)	99 (97 to 100)	⨁⨁⨁⨁ High
False positives (patients incorrectly classified as having drug resistance to pyrazinamide (PZA) (composite))								7 (0 to 21)	5 (0 to 15)	1 (0 to 3)	

Explanations

a. Prevalence of resistance to pyrazinamide (composite) across data used in the model was 56% (CI 52% to 60%)

PICO 2.5: Should tNGS be used to diagnose drug resistance to pyrazinamide (PZA) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Six studies with 425 samples were included in the model for sensitivity, and 6 studies with 379 samples for specificity. Prevalence of resistance to pyrazinamide (pDST) among samples included in the model was 53% (95% CI 49-56%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 14.7% (95% CI 12.3-17.3).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 90% (95% confidence interval (CI) 85%-95%)

Pooled specificity was 90% (95% confidence interval (CI) 86%-94%)

Negative predictive value was 83% (95% confidence interval (CI) 73%-93%)

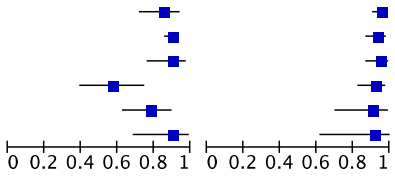
Positive predictive value was 93% (95% confidence interval (CI) 88%-98%)

GRADE assessment:
We downgraded on for inconsistency for sensitivity as the study from Bangladesh was an outlier (see forest plot below). Quality of evidence was thereby assessed to be moderate for sensitivity and high for specificity.

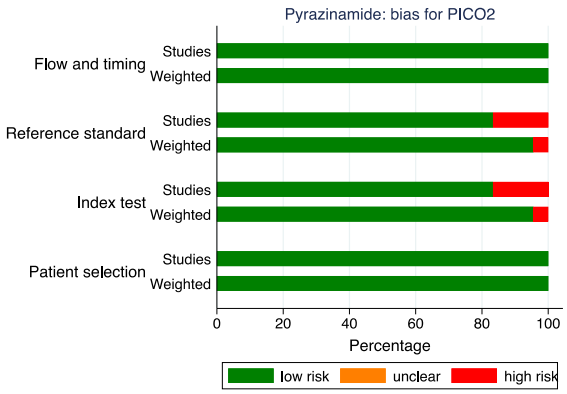
Data by study:

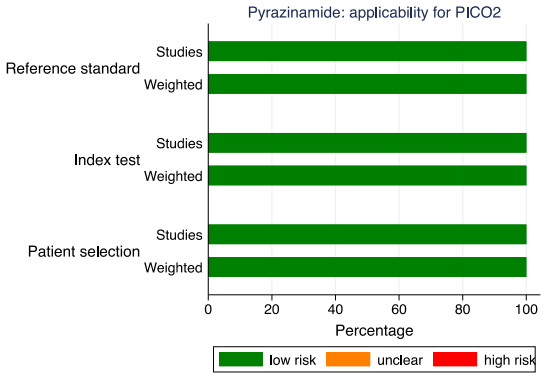
Pyrazinamide

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	40	6	7	123	0.85 [0.72, 0.94]	0.95 [0.90, 0.98]
FIND – India	220	6	25	87	0.90 [0.85, 0.93]	0.94 [0.86, 0.98]
FIND – South Africa	35	3	4	59	0.90 [0.76, 0.97]	0.95 [0.87, 0.99]
icddr,b – Bangladesh	19	5	14	57	0.58 [0.39, 0.75]	0.92 [0.82, 0.97]
IJTLD2022	32	2	9	19	0.78 [0.62, 0.89]	0.90 [0.70, 0.99]
Tuberculosis2021	18	1	2	11	0.90 [0.68, 0.99]	0.92 [0.62, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.90 (95% CI: 0.85 to 0.95)
Specificity	0.90 (95% CI: 0.86 to 0.94)

Prevalences	30%	50%	90%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of30%	pre-test probability of50%	pre-test probability of90%	
True positives (patients with drug resistance to pyrazinamide (PZA) (pDST))	6 studies 425 patients	cross-sectional (cohort type accuracy study)	not serious	not serious	serious[a]	not serious	none	270 (255 to 285)	450 (425 to 475)	810 (765 to 855)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to pyrazinamide (PZA) (pDST))								30 (15 to 45)	50 (25 to 75)	90 (45 to 135)	
True negatives (patients without drug resistance to pyrazinamide (PZA) (pDST))	6 studies 379 patients	cross-sectional (cohort type accuracy study)	not serious	not serious	not serious	not serious	none	630 (602 to 658)	450 (430 to 470)	90 (86 to 94)	⨁⨁⨁⨁ High
False positives (patients incorrectly classified as having drug resistance to pyrazinamide (PZA) (pDST))								70 (42 to 98)	50 (30 to 70)	10 (6 to 14)	

Explanations

a. One study is an outlier for sensitivity

PICO 2.6: Should tNGS be used to diagnose drug resistance to bedaquiline (BDQ) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Three studies with 31 samples were included in the model for sensitivity, and 4 studies with 519 samples for specificity. Prevalence of resistance to bedaquiline among samples included in the model was 6% (95% CI 4-8%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 16.7% (95% CI 13.7-20.1).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 68% (95% confidence interval (CI) 43%-93%)
Model not controlled for semi-quantitative result from MTB/RIF Xpert as this variable was collinear.

Pooled specificity was 97% (95% confidence interval (CI) 94%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear.

Negative predictive value was 99% (95% confidence interval (CI) 99%-100%)
Model not controlled for semi-quantitative result from MTB/RIF Xpert as this variable was collinear.

Positive predictive value was 62% (95% confidence interval (CI) 47%-78%)

GRADE assessment:

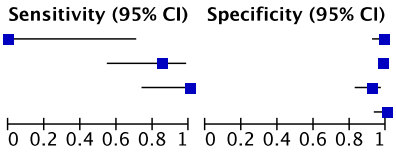
The model for sensitivity was not controlled for semi-quantitative MTB/RIF Xpert results was this variable was collinear in the original model. The model for specificity was not controlled for rifampicin resistance by MTB/RIF Xpert as this variable was collinear in the original model. Instead, the model was restricted to samples that were rifampicin resistant by MTB/RIF Xpert and then controlled for semi-quantitative results. The quality assessment of the evidence was not downgraded as a result.

We downgraded two for imprecision for sensitivity as the numbers of resistant samples were smaller than the threshold we had set for other drugs and as the confidence intervals were very wide, spanning 50 percentage points. Although one of the FIND studies had very low sensitivity, we did not downgrade for inconsistency as the number of resistant samples in that study numbered only 3. Quality of evidence was thereby assessed to be low for sensitivity but high for specificity.

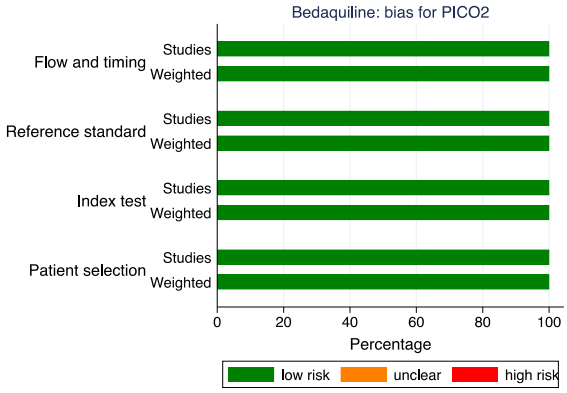
Data by study:

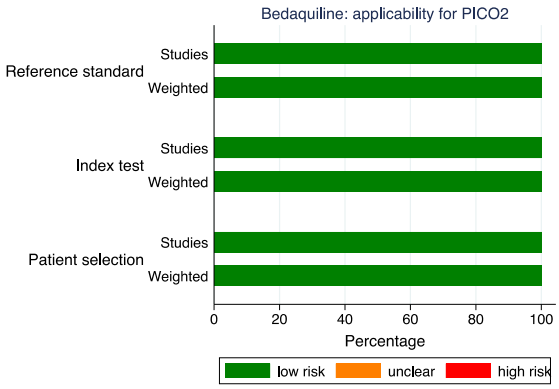
Bedaquiline

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	0	1	3	66	0.00 [0.00, 0.71]	0.99 [0.92, 1.00]
FIND – India	11	7	2	324	0.85 [0.55, 0.98]	0.98 [0.96, 0.99]
FIND – South Africa	12	6	0	65	1.00 [0.74, 1.00]	0.92 [0.83, 0.97]
icddr,b – Bangladesh	0	0	0	50	Not estimable	1.00 [0.93, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.68 (95% CI: 0.43 to 0.93)
Specificity	0.97 (95% CI: 0.94 to 1.00)

Prevalences	1%	3%	5%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of5%	
True positives (patients with drug resistance to bedaquiline (BDQ) (pDST))	3 studies 31 patients[a]	cross-sectional (cohort type accuracy study)	not serious[b]	not serious	not serious[c]	very serious[d]	none	7 (4 to 9)	20 (13 to 28)	34 (22 to 47)	⨁⨁◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to bedaquiline (BDQ) (pDST))								3 (1 to 6)	10 (2 to 17)	16 (3 to 28)	
True negatives (patients without drug resistance to bedaquiline (BDQ) (pDST))	4 studies 519 patients[e]	cross-sectional (cohort type accuracy study)	not serious[b]	not serious	not serious	not serious	none	960 (931 to 990)	941 (912 to 970)	922 (893 to 950)	⨁⨁⨁⨁ High
False positives (patients incorrectly classified as having drug resistance to bedaquiline (BDQ) (pDST))								30 (0 to 59)	29 (0 to 58)	28 (0 to 57)	

Explanations

a. This model is not controlled for CT value as that variable was collinear in the original model
b. Prevalence of resistance to bedaquiline across data used in the model was 6% (CI 4% to 8%)
c. One study had very low sensitivity but it only had 3 resistant samples. It identified 0/3.
d. Very wide 95% confidence intervals for sensitivity
e. This model is not controlled for rifampicin resistance as this variable was collinear in the original model. Instead, the data have been restricted to isolated that are resistant to rifampicin by Xpert, and then controlled for CT value.

PICO 2.7: Should tNGS be used to diagnose drug resistance to linezolid (LZD) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Four studies with 31 samples were included in the model for sensitivity, and 6 studies with 1093 samples for specificity. Prevalence of resistance to linezolid among samples included in the model was 3% (95% CI 2-4%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 15.1% (95% CI 13.1-17.3).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 69% (95% confidence interval (CI) 39%-99%)
Model restricted to samples that were resistant to rifampicin by MTB/RIF Xpert as this variable was collinear in the original model.

Pooled specificity was 100% (95% confidence interval (CI) 100%-100%)
Model restricted to samples that were resistant to rifampicin by MTB/RIF Xpert, and was not controlled for semi-quantitative results as both variables were collinear in the original model.

Negative predictive value was 100% (95% confidence interval (CI) 99%-100%)

Positive predictive value was 93% (95% confidence interval (CI) 84%-100%)
Model not controlled for semi-quantitative result from MTB/RIF Xpert as this variable was collinear, and only includes samples that were resistant to rifampicin by MTB/RIF Xpert as this variable was also collinear in the original model.

GRADE assessment
The model for specificity was restricted to isolates that were resistant to rifampicin by MTB/RIF Xpert and was not controlled for semi-quantitative Xpert result as both results (semi-quantitative and rifampicin resistance were collinear in the original model). The model for sensitivity was also restricted to rifampicin resistant isolates but was controlled for semi-quantitative results as these were not collinear for sensitivity. The quality assessment of the evidence was not downgraded as a result.

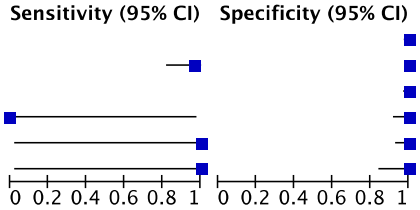
We downgraded two for imprecision for sensitivity as the numbers of resistant samples were smaller than the threshold we had set for other drugs and as the confidence intervals were very wide, spanning 60 percentage points. Although the data from Bangladesh were

an outlier for sensitivity, we did not further downgrade for inconsistency for sensitivity as there was only 1 resistant sample in that study. Quality of evidence was thereby assessed to be low for sensitivity but high for specificity.

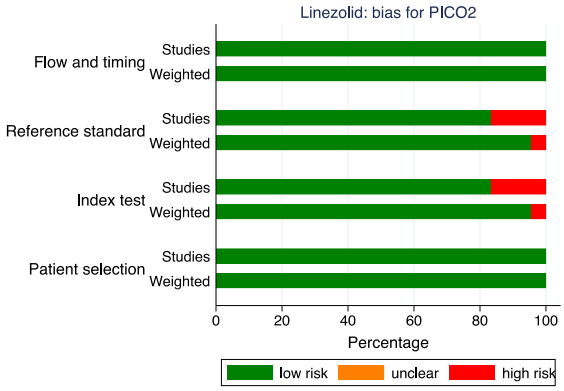
Data by study:

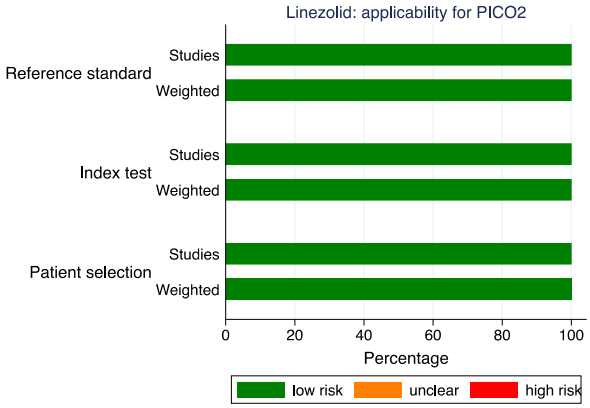
Linezolid

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	0	0	0	123	Not estimable	1.00 [0.97, 1.00]
FIND – India	27	2	1	643	0.96 [0.82, 1.00]	1.00 [0.99, 1.00]
FIND – South Africa	0	0	0	115	Not estimable	1.00 [0.97, 1.00]
icddr,b – Bangladesh	0	0	1	41	0.00 [0.00, 0.97]	1.00 [0.91, 1.00]
IJTLD2022	1	0	0	48	1.00 [0.03, 1.00]	1.00 [0.93, 1.00]
Tuberculosis2021	1	0	0	21	1.00 [0.03, 1.00]	1.00 [0.84, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.69 (95% CI: 0.39 to 0.99)
Specificity	1.00 (95% CI: 1.00 to 1.00)

Prevalences	1%	3%	5%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of5%	
True positives (patients with drug resistance to linezolid (LZD) (pDST))	4 studies 31 patients[a]	cross-sectional (cohort type accuracy study)	not serious[b]	not serious	not serious[c]	very serious[d]	none	7 (4 to 10)	21 (12 to 30)	34 (20 to 50)	⊕⊕◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to linezolid (LZD) (pDST))								3 (0 to 6)	9 (0 to 18)	16 (0 to 30)	
True negatives (patients without drug resistance to linezolid (LZD) (pDST))	6 studies 1093 patients[e]	cross-sectional (cohort type accuracy study)	not serious[b]	not serious	not serious	not serious	none	990 (990 to 990)	970 (970 to 970)	950 (950 to 950)	⊕⊕⊕⊕ High
False positives (patients incorrectly classified as having drug resistance to linezolid (LZD) (pDST))								0 (0 to 0)	0 (0 to 0)	0 (0 to 0)	

Explanations

a. This model is restricted to isolates that were resistant to rifampicin by Xpert, and controls for CT value
b. Prevalence of resistance to linezolid across data used in the model was 3% (CI 2% to 4%)
c. One study was an outlier for sensitivity but only had 1 resistant sample (0/1 detected).
d. Very wide 95% confidence intervals
e. This model is restricted to isolates that were resistant to rifampicin by Xpert, and does not control for CT value as both variables were collinear in the original model

PICO 2.8: Should tNGS be used to diagnose drug resistance to clofazimine (CFZ) (pDST) in patients with bacteriologically confirmed rifampicin-

resistant pulmonary TB disease?

Four studies with 36 samples were included in the model for sensitivity, and 6 studies with 789 samples for specificity. Prevalence of resistance to clofazimine among samples included in the model was 3% (95% CI 2-4%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 11.6% (95% CI 9.5-14.0).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 70% (95% confidence interval (CI) 35%-100%)
Model not controlled for semi-quantitative result from MTB/RIF Xpert as this variable was collinear in the original model.

Pooled specificity was 96% (95% confidence interval (CI) 93%-99%)

Negative predictive value was 99% (95% confidence interval (CI) 98%-100%)
Model not controlled for semi-quantitative result from MTB/RIF Xpert as this variable was collinear.

Positive predictive value was 44% (95% confidence interval (CI) 12%-76%)

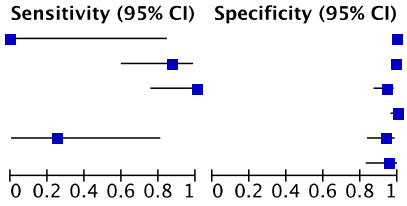
GRADE assessment:
The model for sensitivity was not controlled for semi-quantitative Xpert result as this was collinear in the original model. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for inconsistency for sensitivity two studies were outliers (see forest plot below) and one for imprecision as the confidence intervals were very wide, spanning 65 percentage points. Quality of evidence was thereby assessed to be low for sensitivity but high for specificity.

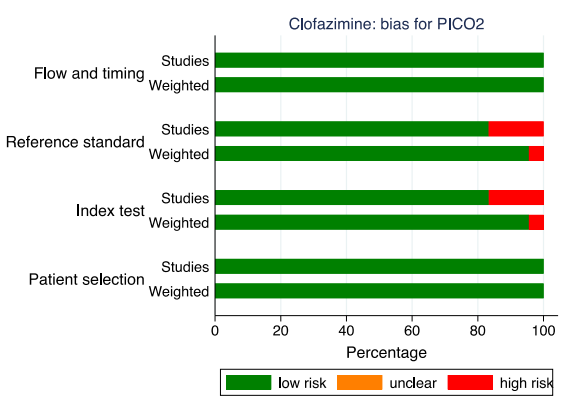
Data by study:

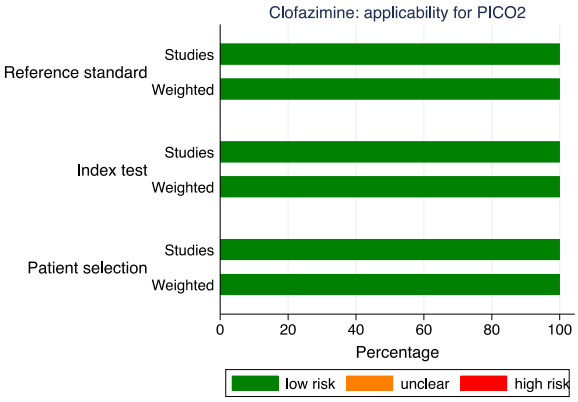
Clofazimine

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	0	1	2	174	0.00 [0.00, 0.84]	0.99 [0.97, 1.00]
FIND – India	13	5	2	329	0.87 [0.60, 0.98]	0.99 [0.97, 1.00]
FIND – South Africa	13	6	0	89	1.00 [0.75, 1.00]	0.94 [0.87, 0.98]
icddr,b – Bangladesh	0	0	0	88	Not estimable	1.00 [0.96, 1.00]
IJTLD2022	1	4	3	54	0.25 [0.01, 0.81]	0.93 [0.83, 0.98]
Tuberculosis2021	0	2	0	37	Not estimable	0.95 [0.83, 0.99]



QUADAS-2 assessment:





Grade table:

Sensitivity	0.70 (95% CI: 0.35 to 1.00)
Specificity	0.96 (95% CI: 0.93 to 0.99)

Prevalences	1%	3%	5%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of1%	pre-test probability of3%	pre-test probability of5%	
True positives (patients with drug resistance to clofazimine (CFZ) (pDST))	4 studies 36 patients[a]	cross-sectional (cohort type accuracy study)	not serious[b]	not serious	serious[c]	serious[d]	none	7 (3 to 10)	21 (10 to 30)	35 (17 to 50)	⨁⨁◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to clofazimine (CFZ) (pDST))								3 (0 to 7)	9 (0 to 20)	15 (0 to 33)	
True negatives (patients without drug resistance to clofazimine (CFZ) (pDST))	6 studies 789 patients	cross-sectional (cohort type accuracy study)	not serious[b]	not serious	not serious	not serious	none	950 (921 to 980)	931 (902 to 960)	912 (884 to 941)	⨁⨁⨁⨁ High
False positives (patients incorrectly classified as having drug resistance to clofazimine (CFZ) (pDST))								40 (10 to 69)	39 (10 to 68)	38 (9 to 66)	

Explanations

a. Model not controlled for CT value as this was collinear in the original model
b. Prevalence of resistance to clofazimine across data used in the model was 3% (CI 2% to 4%)
c. The two smaller studies are outliers for sensitivity. Downgraded as it's more than one small study.
d. Very wide 95% confidence intervals for sensitivity

PICO 2.9: Should tNGS be used to diagnose drug resistance to amikacin (AMK) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Five studies with 115 samples were included in the model for sensitivity, and 8 studies with 1003 samples for specificity. Prevalence of resistance to amikacin among samples included in the model was 10% (95% CI 9-12%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 17.8% (95% CI 15.6-20.2).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 87% (95% confidence interval (CI) 75%-100%)
Model restricted to samples that tested resistant to rifampicin by MTB/RIF Xpert as this variable was collinear in the original model.

Pooled specificity was 99% (95% confidence interval (CI) 98%-100%)
Model restricted to samples that tested resistant to rifampicin by MTB/RIF Xpert as this variable was collinear in the original model.

Negative predictive value was 98% (95% confidence interval (CI) 96%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model

therefore run only on isolates that tested resistant to rifampicin by Xpert.

Positive predictive value was 82% (95% confidence interval (CI) 57%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model therefore run only on isolates that tested resistant to rifampicin by Xpert.

GRADE assessment:

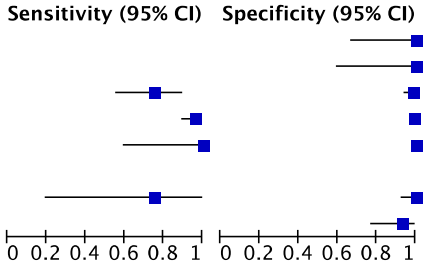
The models for sensitivity and specificity were restricted to samples that were resistant to rifampicin by MTB/RIF Xpert as this variable was collinear in the original model. But both models do control for semi-quantitative Xpert result. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for bias for both sensitivity and specificity as a non WHO-recommended critical concentration was used for a large (almost 40%) proportion of studies. We further downgraded one for inconsistency for sensitivity as there were two outlying studies (see forest plot below) and one for imprecision for sensitivity as the confidence intervals around the point estimate were wide (15). Quality of evidence was thereby assessed to be very low for sensitivity but moderate for specificity.

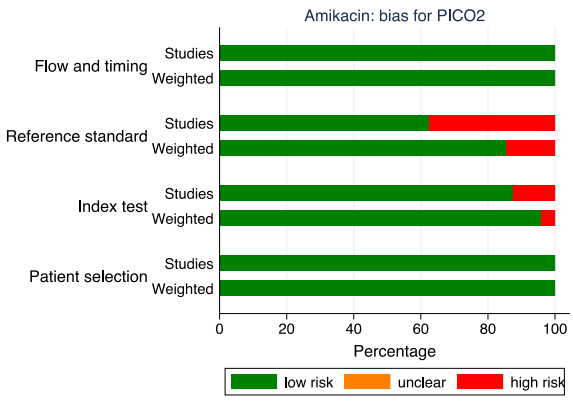
Data by study:

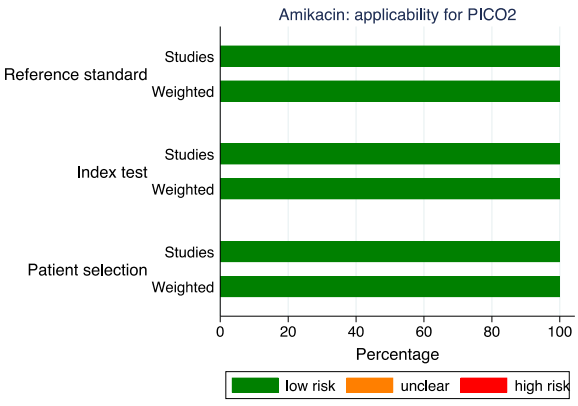
Amikacin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Diama - Benin	0	0	0	9	Not estimable	1.00 [0.66, 1.00]
Diama - Cameroon	0	0	0	7	Not estimable	1.00 [0.59, 1.00]
FIND - Georgia	21	2	7	105	0.75 [0.55, 0.89]	0.98 [0.93, 1.00]
FIND - India	72	6	3	606	0.96 [0.89, 0.99]	0.99 [0.98, 1.00]
FIND - South Africa	7	0	0	150	1.00 [0.59, 1.00]	1.00 [0.98, 1.00]
icddr,b - Bangladesh	0	0	0	0	Not estimable	Not estimable
IJTLD2022	3	0	1	44	0.75 [0.19, 0.99]	1.00 [0.92, 1.00]
Tuberculosis2021	0	2	0	26	Not estimable	0.93 [0.76, 0.99]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.87 (95% CI: 0.75 to 1.00)
Specificity	0.99 (95% CI: 0.98 to 1.00)

Prevalences	5%	10%	15%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of5%	pre-test probability of10%	pre-test probability of15%	
True positives (patients with drug resistance to amikacin (AMK) (pDST))	5 studies 115 patients[a]	cross-sectional (cohort type accuracy study)	serious[b,c]	not serious	serious[d]	serious[e]	none	44 (38 to 50)	87 (75 to 100)	131 (112 to 150)	⨁◯◯◯ Very low
False negatives (patients incorrectly classified as not having drug resistance to amikacin (AMK) (pDST))								6 (0 to 12)	13 (0 to 25)	19 (0 to 38)	
True negatives (patients without drug resistance to amikacin (AMK) (pDST))	8 studies 1003 patients[a]	cross-sectional (cohort type accuracy study)	serious[b,c]	not serious	not serious	not serious	none	941 (931 to 950)	891 (882 to 900)	842 (833 to 850)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having drug resistance to amikacin (AMK) (pDST))								9 (0 to 19)	9 (0 to 18)	8 (0 to 17)	

Explanations

a. The model is restricted to isolated that were resistant to rifampicin by Xpert, as this was collinear in the original model, but controls for CT value
b. Prevalence of resistance to amikacin across data used in the model was 10% (CI 9% to 12%)
c. Non WHO recommended CC used
d. Two outlying studies for sensitivity, albeit small studies
e. wide 95% confidence intervals for sensitivity

PICO 2.10: Should tNGS be used to diagnose drug resistance to ethambutol (EMB) (composite) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Four studies with 431 samples were included in the model for sensitivity, and 4 studies with 123 samples for specificity. Prevalence of resistance to ethambutol (composite) among samples included in the model was 78% (95% CI 74-81%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 20.6% (95% CI 17.3-24.2).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 97% (95% confidence interval (CI) 95%-98%)

Pooled specificity was 98% (95% confidence interval (CI) 96%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model therefore run only on isolates that tested resistant to rifampicin by Xpert.

Negative predictive value was 89% (95% confidence interval (CI) 81%-96%)

Positive predictive value was 100% (95% confidence interval (CI) 99%-100%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model therefore run only on isolates that tested resistant to rifampicin by Xpert.

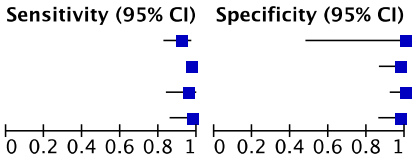
GRADE assessment:
The model for specificity was restricted to samples that were resistant to rifampicin by MTB/RIF Xpert as this variable was collinear in the original model. The model does control for semi-quantitative Xpert result. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for bias for both sensitivity and specificity as different samples were tested by the index and reference tests. Quality of evidence was thereby assessed to be moderate for both sensitivity and specificity.

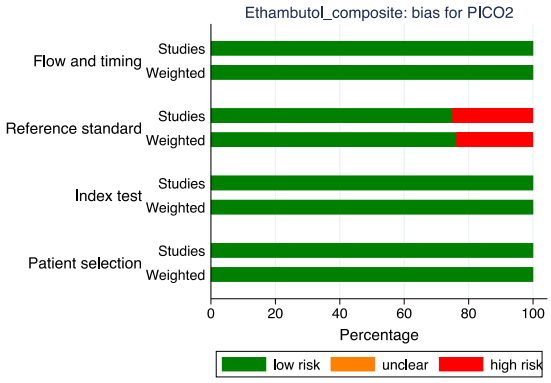
Data by study:

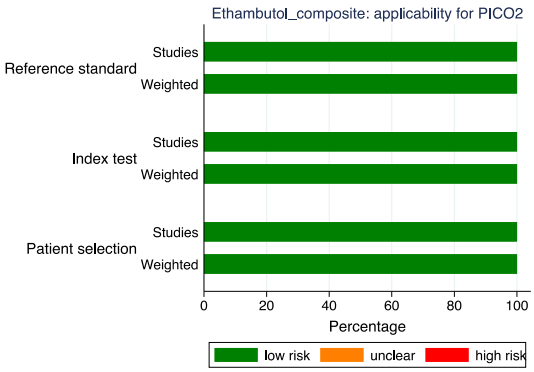
Ethambutol_composite

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	64	0	6	5	0.91 [0.82, 0.97]	1.00 [0.48, 1.00]
FIND – India	275	1	9	37	0.97 [0.94, 0.99]	0.97 [0.86, 1.00]
FIND – South Africa	39	0	2	43	0.95 [0.83, 0.99]	1.00 [0.92, 1.00]
TBSeq – China	35	1	1	36	0.97 [0.85, 1.00]	0.97 [0.86, 1.00]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.97 (95% CI: 0.95 to 0.98)
Specificity	0.98 (95% CI: 0.96 to 1.00)

Prevalences	10%	30%	50%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of10%	pre-test probability of30%	pre-test probability of50%	
True positives (patients with drug resistance to ethambutol (EMB) (composite))	4 studies 431 patients	cross-sectional (cohort type accuracy study)	serious[a,b]	not serious	not serious	not serious	none	97 (95 to 98)	291 (285 to 294)	485 (475 to 490)	⨁⨁⨁◯ Moderate
False negatives (patients incorrectly classified as not having drug resistance to ethambutol (EMB) (composite))								3 (2 to 5)	9 (6 to 15)	15 (10 to 25)	
True negatives (patients without drug resistance to ethambutol (EMB) (composite))	4 studies 123 patients[c]	cross-sectional (cohort type accuracy study)	serious[a,b]	not serious	not serious	not serious	none	882 (864 to 900)	686 (672 to 700)	490 (480 to 500)	⨁⨁⨁◯ Moderate
False positives (patients incorrectly classified as having drug resistance to ethambutol (EMB) (composite))								18 (0 to 36)	14 (0 to 28)	10 (0 to 20)	

Explanations

a. Prevalence of resistance to ethambutol (composite) across data used in the model was 78% (CI 74% to 81%)
b. Different samples tested for index and reference tests
c. The model is restricted to isolated that were resistant to rifampicin by Xpert, as this was collinear in the original model, but controls for CT value

PICO 2.11: Should tNGS be used to diagnose drug resistance to ethambutol (EMB) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

One study with 89 samples was included in the model for sensitivity, and 1 study with 213 samples for specificity. Prevalence of resistance to ethambutol (pDST) among samples included in the model was 29% (95% CI 24-35%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 0% (95% CI 0-1).

Results from multivariable, mixed-effects model:

Pooled sensitivity was 91% (95% confidence interval (CI) 85%-97%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model

therefore run only on isolates that tested resistant to rifampicin by Xpert.

Pooled specificity was 92% (95% confidence interval (CI) 88%-96%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model therefore run only on isolates that tested resistant to rifampicin by Xpert.

Negative predictive value was 96% (95% confidence interval (CI) 93%-99%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model therefore run only on isolates that tested resistant to rifampicin by Xpert.

Positive predictive value was 83% (95% confidence interval (CI) 75%-90%)
Model not controlled for rifampicin resistance as determined by MTB/RIF Xpert as this variable was collinear. Model therefore run only on isolates that tested resistant to rifampicin by Xpert.

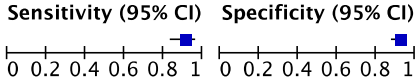
GRADE assessment:
The models for both sensitivity and specificity were restricted to samples that were resistant to rifampicin by MTB/RIF Xpert as this variable was collinear in both original models. The models do however control for semi-quantitative Xpert results. The quality assessment of the evidence was not downgraded as a result.

We downgraded one for bias for both sensitivity and specificity as different samples were tested by the index and reference tests. We further downgraded one for indirectness for each of sensitivity and specificity as the data were only from one country and there could be concerns around generalizability. Quality of evidence was thereby assessed to be low for both sensitivity and specificity.

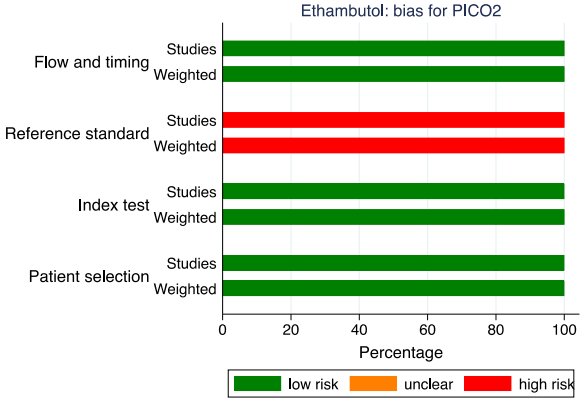
Data by study:

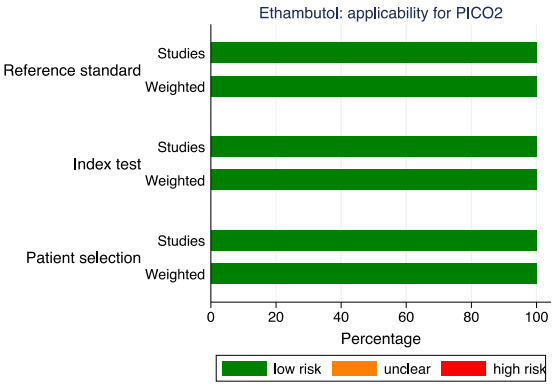
Ethambutol

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
TBSeq - China	81	17	8	196	0.91 [0.83, 0.96]	0.92 [0.88, 0.95]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.91 (95% CI: 0.85 to 0.97)
Specificity	0.92 (95% CI: 0.88 to 0.96)

Prevalences	10%	30%	50%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of10%	pre-test probability of30%	pre-test probability of50%	
True positives (patients with drug resistance to ethambutol (EMB) (pDST))	1 studies 89 patients[a]	cross-sectional (cohort type accuracy study)	serious[b,c]	serious[d]	not serious	not serious	none	91 (85 to 97)	273 (255 to 291)	455 (425 to 485)	⨁⨁◯◯ Low
False negatives (patients incorrectly classified as not having drug resistance to ethambutol (EMB) (pDST))								9 (3 to 15)	27 (9 to 45)	45 (15 to 75)	
True negatives (patients without drug resistance to ethambutol (EMB) (pDST))	1 studies 213 patients[a]	cross-sectional (cohort type accuracy study)	serious[b,c]	serious[d]	not serious	not serious	none	828 (792 to 864)	644 (616 to 672)	460 (440 to 480)	⨁⨁◯◯ Low
False positives (patients incorrectly classified as having drug resistance to ethambutol (EMB) (pDST))								72 (36 to 108)	56 (28 to 84)	40 (20 to 60)	

Explanations

a. The model is restricted to isolated that were resistant to rifampicin by Xpert, as this was collinear in the original model, but controls for CT value
b. Different samples used for tNGS and reference test
c. Prevalence of resistance to ethambutol (pDST) across data used in the model was 29% (CI 24% to 35%)
d. Only one study (from China). Downgraded by one as may not be generalisable.

PICO 2.12: Should tNGS be used to diagnose drug resistance to streptomycin (STR) (pDST) in patients with bacteriologically confirmed rifampicin-resistant pulmonary TB disease?

Five studies with 493 samples were included in the model for sensitivity, and 5 studies with 250 samples for specificity. Prevalence of resistance to streptomycin among samples included in the model was 66% (95% CI 63-70%). The proportion of samples with no tNGS result reported (the 'indeterminate rate') was 18.8% (95% CI 16.1-21.8).

Results from multivariable, mixed-effects model:
Pooled sensitivity was 98% (95% confidence interval (CI) 96%-100%)

Pooled specificity was 75% (95% confidence interval (CI) 59%-91%)

Negative predictive value was 91% (95% confidence interval (CI) 82%-100%)

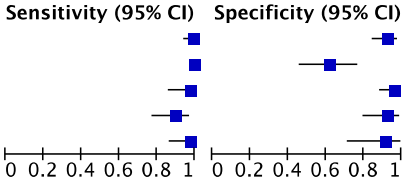
Positive predictive value was 95% (95% confidence interval (CI) 93%-97%)

GRADE assessment:
We downgraded one for inconsistency for specificity as one study was an outlier (see forest plot below). We further downgraded one for imprecision for specificity as the confidence intervals around the point estimate were wide (32 percentage points). Quality of evidence was thereby assessed to be high for sensitivity but low for specificity.

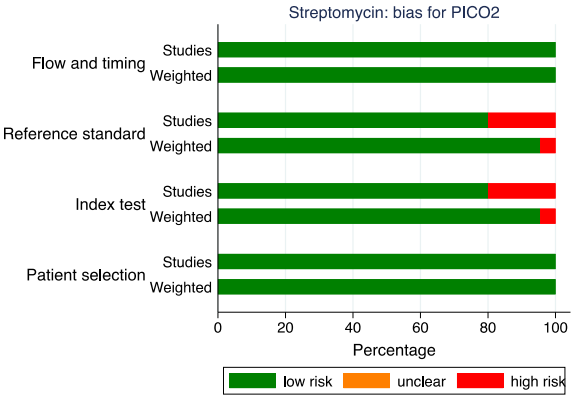
Data by study:

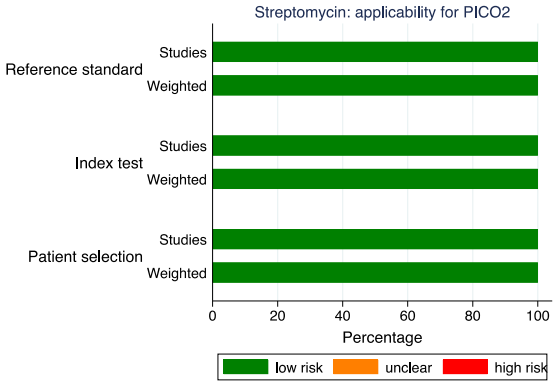
Streptomycin

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
FIND – Georgia	81	6	1	72	0.99 [0.93, 1.00]	0.92 [0.84, 0.97]
FIND – India	288	16	3	26	0.99 [0.97, 1.00]	0.62 [0.46, 0.76]
FIND – South Africa	35	3	1	66	0.97 [0.85, 1.00]	0.96 [0.88, 0.99]
icddr,b – Bangladesh	42	3	5	36	0.89 [0.77, 0.96]	0.92 [0.79, 0.98]
IJTLD2022	36	2	1	20	0.97 [0.86, 1.00]	0.91 [0.71, 0.99]



QUADAS-2 assessment:





GRADE table:

Sensitivity	0.98 (95% CI: 0.96 to 1.00)
Specificity	0.75 (95% CI: 0.59 to 0.91)

Prevalences	10%	30%	50%

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 patients tested			Test accuracy CoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of10%	pre-test probability of30%	pre-test probability of50%	
True positives (patients with drug resistance to streptomycin (STR) (pDST))	5 studies 493 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	not serious	not serious	none	98 (96 to 100)	294 (288 to 300)	490 (480 to 500)	⊕⊕⊕⊕ High
False negatives (patients incorrectly classified as not having drug resistance to streptomycin (STR) (pDST))								2 (0 to 4)	6 (0 to 12)	10 (0 to 20)	
True negatives (patients without drug resistance to streptomycin (STR) (pDST))	5 studies 250 patients	cross-sectional (cohort type accuracy study)	not serious[a]	not serious	serious[b]	serious[c]	none	675 (531 to 819)	525 (413 to 637)	375 (295 to 455)	⊕⊕◯◯ Low
False positives (patients incorrectly classified as having drug resistance to streptomycin (STR) (pDST))								225 (81 to 369)	175 (63 to 287)	125 (45 to 205)	

Explanations

a. Prevalence of resistance to streptomycin across data used in the model was 66% (CI 63% to 70%)
b. One study was an outlier
c. Wide 95% confidence intervals for specificity

Author's conclusions

Regarding positioning tNGS as the first test for drug resistance for patients (PICO 1):

Pooled sensitivity to first line drugs was 88%-96%, with the highest estimate for isoniazid and the lowest for pyrazinamide. This is similar to the results obtained from WGS sequencing from culture,(10) a routine diagnostic approach already adopted by a number of countries. Sensitivity for the flouroquinolones was 94%-96%, which is comparable or higher than some results obtained from WGS from culture. This may be because tNGS better detects low frequency resistance alleles.(14)

Pooled specificity to first line drugs was 96%-99%, with the highest estimate for ethambutol and the lowest for rifampicin. This is again comparable to the results obtained from WGS sequencing from culture. Specificity for the flouroquinolones was 96%, indicating minimal loss in specificity for any gain in sensitivity through the detection of low frequency resistance alleles.

Any recommendation to position tNGS as the first test of drug resistance needs to consider a wide range of factors, including the added value of diagnosing isoniazid mono-resistance which is not detected by MTB/RIF Xpert® or Xpert Ultra®, although is detected by Xpert XDR®. A head-to-head study would help inform such a decision, alongside other considerations including cost and operability.

Regarding positioning tNGS as a reflex test for patients with samples known to be

resistant to rifampicin (PICO 2):

Pooled sensitivity to all of the drugs included in PICO 1 above is marginally higher whereas pooled specificity is marginally lower with the exception of pyrazinamide. The sensitivity for bedaquiline was 68%, for clofazimine 70%, and for linezolid 69%, with specificity for all three over 96%. These figures are clearly lower than for the legacy drugs but given that there is currently no molecular DST for these drugs, and given that the WHO catalogue of mutations in *M. tuberculosis* that are associated with drug resistance included only one resistance mutation for linezolid and none for bedaquiline or clofazimine in 2021, these are encouraging results.(15)

The quality of evidence was assessed as being lower, on the whole, for PICO 1 than for PICO 2. However, the evidence was downgraded in accordance with the rules of the GRADE method, and in the opinion of the authors, the evidence presented in this review nevertheless indicates that the best performing platforms within the class of tNGS assays deliver impressive accuracy. What is however of concern is the indeterminate rate that is between 10-20% for most drugs. This is a limitation and the extent to which this will improve remains to be shown. Future studies will be needed to assess diagnostic accuracy for extra-pulmonary specimens.

References

1. Global Tuberculosis Report 2022. Geneva: World Health Organization; Licence: CC BY-NC-SA 3.0 IGO

2. Karmakar M, Ragonnet R, Ascher DB, Trauer JM, Denholm JT. Estimating tuberculosis drug resistance amplification rates in high-burden settings. BMC Infect Dis. 2022 Jan 24;22(1):82.

3. The End TB Strategy. Geneva: World Health Organization; WHO/HTM/TB/2015.19. Available from: https://www.who.int/publications/i/item/WHO-HTM-TB-2015.19 (Accessed 6 April 2023)

4. WHO consolidated guidelines on tuberculosis. Module 3: diagnosis - rapid diagnostics for tuberculosis detection, 2021 update. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO

5. Dartois VA, Rubin EJ. Anti-tuberculosis treatment strategies and drug development: challenges and priorities. Nat Rev Microbiol. 2022 Nov;20(11):685–701.

6. Dorman SE, Nahid P, Kurbatova EV, Phillips PPJ, Bryant K, Dooley KE, et al. Four-Month Rifapentine Regimens with or without Moxifloxacin for Tuberculosis. N Engl J Med. 2021 May 6;384(18):1705–18.

7. Rapid communication: key changes to the treatment of drug-resistant tuberculosis. Geneva: World Health Organization; 2022 (WHO/UCN/TB/2022.2). Licence: CC BY-NC-SA 3.0 IGO

8. Lindsay McKenna. The 2021 Tuberculosis Treatment Pipeline Report. Treatment Action Group. Available from: https://www.treatmentactiongroup.org/wp-content/uploads/2021/11/pipeline_TB_Treatment_2021_final.pdf (Accessed 16 April 2023)

9. Van Rie A, Walker T, de Jong B, Rupasinghe P, Rivière E, Dartois V, et al. Balancing access to BPaLM regimens and risk of resistance. The Lancet Infectious Diseases. 2022 Oct;22(10):1411–2.

10. The CRyPTIC Consortium and the 100 000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018 Sep 26;NEJMoa1800474-14.

11. Dale K, Globan M, Horan K, Sherry N, Ballard S, Tay EL, et al. Whole genome sequencing for tuberculosis in Victoria, Australia: A genomic implementation study from 2017 to 2020. The Lancet Regional Health - Western Pacific. 2022 Nov;28:100556.

12. Nilgiriwala K, Rabodoarivelo MS, Hall MB, Patel G, Mandal A, Mishra S, et al. Genomic Sequencing from Sputum for Tuberculosis Disease Diagnosis, Lineage Determination, and Drug Susceptibility Prediction. Diekema DJ, editor. J Clin Microbiol. 2023 Mar 23;61(3):e01578-22.

13. Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011 Oct 18;155(8):529.

14. Brankin AE, Fowler PW. Inclusion of minor alleles improves catalogue-based prediction of
fluoroquinolone resistance in Mycobacterium tuberculosis. JAC Antimicrob Resist 2023 Apr
4;5(2)

15. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO

Web Annex D.19. A Systematic and Scoping Review of the Economic Evidence around Targeted Next Generation Sequencing (tNGS) for the Diagnosis of Drug-Resistant Tuberculosis (DR-TB)

Prepared for WHO Global TB Programme, Guideline Development Group

Patrick Galange[1], Suvesh Shrestha[2], Israa Zareef[2], Alice Zwerling[2]
[1] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[2] School of Epidemiology & Public Health, University of Ottawa, Ottawa, Canada

Corresponding Author:
Alice Zwerling
azwerlin@uottawa.ca

EXECUTIVE SUMMARY

Background: Tuberculosis (TB) remains one of the leading causes of morbidity and mortality globally. Our review aims to summarize the current economic evidence around using targeted Next Generation Sequencing (tNGS) for the diagnosis of DR-TB.
Methods: We conducted a systematic review on the economic evaluation of using either tNGS or WGS to diagnose DR-TB. The following three databases were used: PubMed, EMBASE and SCOPUS. The search was run on October 30, 2022. We did not restrict by year of publication; age group; country; income-level; comparator group; HIV status, or other comorbidities. All costing data were inflated to 2021 USD. Findings were synthesized descriptively given the considerable degree of heterogeneity in study methodology and outcomes. A scoping review was also performed to add data from other disease areas.
Results: Overall, there were 10 studies included in our systematic review, which assessed tNGS only (n=3), tNGS and WGS (n=3), or WGS only (n=4). For tNGS (n=1), the cost per sample was between $69.64 for Illumina MiSeq on 24 samples, and $73.47 for Nanopore MinION on 12 samples. For WGS (n=5), cost per sample ranged from $63.00 on Nanopore MinION to $277.00 on Illumina MiSeq. The step with the greatest cost was sequencing, and the most significant component costs were reagents and consumables. There were four major cost drivers identified by authors: using different sequencers, depth and breadth of coverage, inefficiencies in initial sample runs, and economies of scale via batching or cross-batching. The scoping review corroborated the importance of which sequencer was being used, reagents and consumables being the most significant component cost, sequencing being the step with the greatest cost, as well as similar cost drivers. In addition, there was a 2001 cost-effectiveness analysis included, and a number of additional cost drivers identified. This includes the operational efficiency of the lab, availability of trained personnel, sequencers being used at full capacity, discounts associated with purchasing high volume from the same suppliers, and complexity of infectious pathogen.
Conclusions: Our systematic and scoping reviews are the first of their kind to assess the economic evidence around tNGS and WGS for the diagnosis of DR-TB. Further cost-effectiveness analyses and more in-depth costing data in this space would be helpful to better understand the potential for tNGS relative to existing drug sensitivity testing, and to guide future scale-up decisions.

1. INTRODUCTION

Tuberculosis remains one of the leading causes of morbidity and mortality globally (1). In 2021, the World Health Organization (WHO) estimated approximately 10.6 million new cases of TB occurred, and over 1.6 million deaths, the latter being an increase in mortality compared to trends seen between 2005 and 2019 (2, 3). DR-TB was first recognized in the late 1940s and has become an increasing public health concern (2). DR-TB can be further classified as Rifampin-resistant TB (RR-TB), Isoniazid-resistant TB (IR-TB), Multidrug-resistant TB (MDR-TB), Pre-extensively Drug-resistant TB (Pre-XDR-TB) and Extensively drug-resistant TB (XDR-TB) (3, 4). These groupings refer to *Mycobacterium tuberculosis* with resistance to rifampicin alone; isoniazid alone; at least isoniazid and rifampicin; rifampicin and any fluoroquinolone; and rifampicin, a fluoroquinolone as well as at least one either bedaquiline or linezolid; respectively (3, 4). In 2015, it was estimated that 3.9% of new TB diagnoses, and 21% of previously-treated TB cases were either rifampicin- or multidrug-resistant globally, however these estimates vary widely across different settings (5). In 2020, the 30 high burden TB countries accounted for 87% of new cases, with more than two thirds of cases occurring in eight countries (6). This includes India, Indonesia, China, the Philippines, Pakistan, Nigeria, Bangladesh, and the Democratic Republic of the Congo (6). When considering which countries are the highest need to introduce improved testing for DR-TB, it is not only the absolute number, but the proportion of new TB cases with any drug resistance that matters. Indeed, the proportion of new TB cases with any drug resistance was also disproportionately higher in certain countries. For example, the global incidence of MDR/RR-TB is 5.7 per 100,000 population, but as high as 8.4 per 100,000 in the WHO South-East Asia region and as low as 1.2 per 100,000 in the WHO/PAHO Region of the Americas (3, 7). Unfortunately, the incidence of DR-TB was estimated to have increased by 3.1% in 2021, with 450,000 newly diagnosed with DR-TB, cases in 2021 according to the most recent Global Tuberculosis Report (3).

Due to the extent of the global DR-TB problem and the WHO's End TB Strategy by 2030, there has been significant attention and growth in tuberculosis diagnostics aimed at the detection of DR (8). However, it has remained a challenge to develop an accurate, rapid, affordable, and accessible method for drug-susceptibility testing (DST). At present, there exists several strategies endorsed by the WHO that can be used for the diagnosis of DR-TB (9). These can be broadly divided into either molecular or phenotypic testing (10). Probe-based molecular testing, such as GeneXpert (Cepheid Inc., Sunnyvale, CA, USA) or TrueNAT (Molbio Diagnostics, India), are the first-line approaches in most countries (11, 12). These PCR-based tests work through identifying known resistance conveying mutations in specific genes, and together are referred to as molecular WHO-recommended rapid diagnostics (mWRDs) (11). Benefits include rapid turnaround (as little as two hours); minimal specialized lab training requirements; and the ability of mWRDs to diagnose RR-TB, which is a reliable marker for MDR-TB (13, 14). Sensitivity (97.6%) and specificity (99.2%) of GeneXpert are high and these tests are often used as the primary diagnostic test in many countries for individuals with TB symptoms. Once MDR-TB or RR-TB is diagnosed via mWRDs, the next steps are often country-dependent. In many places, this would involve treatment with an MDR-TB drug regimen, such as an abbreviated bedaquiline-containing regiment for six months (13). If the patient fails treatment, they may be referred for additional DST depending on availability; as mWRDs are limited in that they are only able to accurately diagnose mutations within certain known regions and thereby limited in their ability to detect a range of resistance profiles (15).

Second-line drug sensitivity testing using conventional phenotypic DST (pDST) relies on either liquid or solid cultures to confirm *Mycobacterium tuberculosis* and drug resistance by measuring growth and minimum inhibitory concentrations (MIC) in the presence of a given anti-TB drug (10, 13, 16, and 17). This approach can be lengthy requiring a minimum of several weeks for sufficient bacterial growth and the need for biosafety level 3 laboratories and highly trained staff leading to

delays in diagnosis and treatment, negatively impacting patient outcomes (17, 18).

Alternative second-line diagnostic approaches for DR include Line Probe Assays (LPAs), which avoid many of the steps in conventional pDST (19). This is another type of probe-based molecular test. In 2008, the WHO approved the use of LPAs for the diagnosis of MDR-TB, and there are currently ten different types commercially available (19, 20). Most LPAs can only diagnose rifampin resistance via the *rpoB* gene, and/or isoniazid resistance via the *katG* gene; but newer modalities can also detect resistance in ethambutol, fluoroquinolones, and injectable agents via the *embB*, *gyrA/gyrB,* and *rrs* genes, respectively (21). The advantages of LPAs are that they are highly effective at diagnosing tuberculosis, rifampicin resistance, isoniazid resistance, and MDR-TB; and they can detect samples within only a few hours by avoiding the need for culture (20). However, LPAs have low sensitivity for pre-XDR-TB, XDR-TB, and extra-pulmonary tuberculosis; are quite laboratory and labor intensive; can be expensive; and are only sensitive in cases where the sputum are positive. As a result, LPAs are recommended by the WHO as being more suitable as a complementary method in middle-incidence countries that have more laboratory access (20).

Recently more development and interest has been seen in the area of sequencing-based approaches for diagnosing drug resistance in TB (22). Next Generation Sequencing refers to a "high-throughput, massively parallel" sequencing technology that uses a single biochemical reaction to determine a genetic sequence (23). This can be divided into WGS, which looks at the entire genome; tNGS, which looks at part of one; or metagenomics, which is a culture-independent analysis of multiple microbes from the environment (24, 25). End-to-end sequencing solutions for DR-TB diagnosis typically involves the following steps: DNA extraction, library preparation, targeted sequencing, data analysis & interpretation, and the generation of a final report (26). Compared to WGS, tNGS has the potential to be more rapid, scalable, and cost-effective; and as a result, has garnered much attention as a promising new approach for the diagnosis of DR-TB (27). The routine use of tNGS has been made possible in recent years for many high-income countries (HICs) and some LMICs due to improvements in cost and operational efficiency (28). The advantages of tNGS for DR-TB are broad, including improved speed (48 hours versus 6-8 weeks for conventional pDST); as well as comprehensive coverage of many more potential mutations (29). Despite this, there have been barriers to the adoption of tNGS for DST in LMICs, which include: 1) perceived technical complexity and high cost, 2) lack of an end-to-end commercial solution, and 3) lack of real-world data around patient important outcomes to inform policy and update guidelines (30).

The primary objective of this study was to perform a systematic review of the available economic evidence around tNGS for the diagnosis of DR-TB, not limited to one setting or subpopulation. Due to the anticipated paucity of studies, a subsequent complimentary scoping review was planned to assess economic evidence for tNGS in a number of other infectious pathogens, including: HIV, hepatitis B virus (HBV), hepatitis C virus (HCV), malaria and influenza (31-34).

3. METHODS

2.1. Data Collection

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for conducting and reporting systematic reviews (35). Our search strategy was developed with the support of a Johns Hopkins University Welsch librarian, and is shown in Appendix 8.2. For the systematic review, the search strategy contained three overarching concepts: concept 1 (TB and

DR-TB); concept 2 (economic data, such as cost or cost-effectiveness); and concept 3 (sequencing technologies, such as tNGS).

We used the following three electronic databases to perform the search as of October 30, 2022: PubMed, EMBASE, and SCOPUS for new studies, not restricted by year of publication or language. Additionally, we screened the reference list of review articles and key identified articles for additional studies.

For the systematic review, we included studies that looked at participants with a confirmed or suspected diagnosis of pulmonary or non-pulmonary DR-TB via any WHO recommended phenotypic or molecular assay. Studies that included only DS-TB participants were excluded. For both reviews, the diagnostic intervention of interest was either tNGS or WGS. Even though the diagnostic intervention of primary interest was tNGS, we elected to include studies that looked at WGS given the known overlap between the lab methodologies of each.

Studies were not excluded based on age group; country; income-level; HIV status, or other comorbidities; sample type (sputum versus non-sputum); test manufacturer; or by the presence or absence of a comparator. Finally, we did not exclude any study designs, or type of epidemiologic or economic outcome, including costing studies, cost-effectiveness analyses, and cost-benefit analyses.

Titles and abstracts retrieved with the aforementioned search strategy were screened by two independent reviewers, PG and SS, as per our inclusion and exclusion criteria. A full text review was then performed by two reviewers to independently assess studies for eligibility. All disagreements were resolved through discussion, a third reviewer (AZ) was consulted when a consensus was not reached (35). This same process was followed for data extraction.

2.2. Data Extraction

Data was extracted from the included studies for assessment of study characteristics, study methodologies, data output, and quality of evidence (35). A standardized set of extraction sheets were generated through an iterative process, based on key data of interest for both reviews. For the systematic review, key study characteristics included country setting, study population, diagnostic strategies (tNGS or WGS), comparators if included, economic analysis perspective (healthcare or societal), type of economic evaluation (cost analysis or other), year of cost valuation, currency, primary outcome, secondary outcome(s), and source of costing. Key methodological elements extracted in the systematic review included different scenarios modelled for costing, key scenarios/variables explored in sensitivity analyses, sample run turnover time, depth of sequencing coverage for costing, as well as breadth of sequencing coverage for costing. Depth of coverage refers to the number of times one examines a single base in question; whereas breadth of coverage refers to the proportion of the genome in WGS or targets in tNGS that are covered by sequencing data.

2.3. Outcomes

Key outcomes of interest in the included tNGS studies were: unit test costs or cost per patient for each diagnostic strategy, including cost per patient tested and cost per patient diagnosed; any component cost associated with the NGS technology or implementation, such as overhead, labor, or equipment; any cost associated with a specific step of the NGS pathway, such as library preparation, DNA replication, or sequencing; and incremental cost effectiveness ratio (ICER) per effectiveness or utility measure (e.g., case detected, DALY averted or QALY gained). For WGS, we identified areas where costs would likely be comparable to that of tNGS, such as key component costs including overhead, or labor costs; or costs associated with sample collection.

2.4. Statistical Analysis

Due to the limited number of studies available and the high degree of methodological heterogeneity between studies, a meta-analysis was deemed not appropriate. Instead, a narrative summary review was performed.

All cost data was inflated to 2021 in local currency and converted to 2021 USD using the relevant International Monetary Fund (IMF) exchange rates and inflation index (36). We contacted study authors for missing or unclear data.

2.5. Assessing the Quality of Evidence

We used a modified version of the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement to assess the quality of evidence for each study (37). The CHEERS statement aims to consolidate and update previous health economic evaluation guidelines into a single and up-to-date guidance (37). We elected to use a modified version of CHEERS because there was a paucity of cost-effectiveness studies. As a result, many of the CHEERS components were not relevant. The full list of included components is shown in the Appendix (Supplementary Tables 1A-3B).

2.6. Scoping Review

The methodology of the scoping review was similar to that of the systematic review, except rather than diagnosing DR-TB, we expanded the search to include comparable infectious diseases (Appendix 8.3). These included one or more of the following: HIV, malaria, HBV, HCV, and influenza. This led to some variations in the extraction sheets, such as including identity of communicable disease in study characteristics (Tables 3A-B). Of note, one of the included studies considered avian influenza, influenza A&B, as well as food-borne pathogens; and another study considered multiple avian paramyxoviruses. We decided to include these because they had relevant data, and there was a paucity of manuscripts meeting our inclusion criteria.

3.RESULTS

3.1. Systematic Review

In total, there were 79 studies identified by our search strategy after removing duplications. Of these, 10 were included in the final systematic review (Figure 1) (16, 38-46). Articles were excluded if they did not include persons with TB, did not use either tNGS or WGS as a diagnostic intervention, or did not include any economic or cost data. Of the 10 selected articles, 1 did bottom-up costing through micro-costing, and 1 included a budget impact analysis (BIA). Three studies explored costs exclusively for tNGS, 4 explored costs exclusively for WGS, and 3 explored costs for both tNGS and WGS. There were no direct comparisons using a cost-effectiveness analysis, and no studies reported incremental cost effectiveness ratios (ICER) for either tNGS or WGS. Included tNGS studies reported on diagnostic cost per sample; the component costs equipment and sequencers, reagents and consumables, labor and training, and maintenance; as well as costs per step, including target enrichment multiplex PCR, library prep, quality control, sequencing, and DNA extraction. Included WGS studies reported on diagnostic cost per sample; the component costs including equipment and sequencers, reagents and consumables, labor and training, overhead, and maintenance; as well as costs per step, including target enrichment multiplex PCR, library prep, quality control, sequencing, and DNA extraction. Only one of the tNGS, and two of the WGS papers assessed costs associated with the diagnostic comparator pDST.

3.1.1. Economic Evidence Around tNGS

3.1.1.1. Summary of Findings

Six studies reported on costs for tNGS approaches (Table 1). Three of the studies were performed in the US, one collected specimen from both Ethiopia and Hong Kong, one in the UK, and one in Moldova with year of cost valuation ranging from 2018-2021 (Table 1). Of the tNGS studies, two used on-site testing, and four required sample transport to a central lab for processing (Table 1). We identified two studies that assessed diagnostic cost per sample for tNGS, and one that assessed this for pDST as a diagnostic comparator. Five of the manuscripts considered component costs for tNGS, and one manuscript considered component costs for pDST. The manuscript by Tafess K *et al.* (2020) was the only one to consider cost per step of tNGS (38). The study by Cates *et al.* (2022) was the only manuscript to perform a BIA for tNGS (39).

3.1.1.2. Diagnostic Cost per Sample

For tNGS, the diagnostic cost per sample ranged from $69.64 to $73.47 (n=1). The lower value was run on Illumina MiSeq with 24 samples/run, top-down costing, 15,000-fold coverage, and including some equipment and reagents and consumables costs but with no labour and overhead costs (Table 7A). The upper value was run on Nanopore MinION with 12 samples/run, top-down costing, 15,000-fold coverage, and including some equipment and reagents and consumables costs but with no labour and overhead costs (Table 7A). One key influential factor on costs was which sequencing platform was used, with Illumina iSeq100 having the highest costs, followed by Nanopore MinION, and then Illumina MiSeq. However, it should be noted that the above findings are based on limited data. For example, it was not clear whether the authors accounted for the fact that the Nanopore sequencer is only available through being rented, nor the added costs for required computational equipment.

3.1.1.3. Costs per Step

The paper by Tafess *et al* (2020) was the only one to break down diagnostic cost per sample into cost per step for each sample (38). These steps included target enrichment multiplex PCR, library prep, quality control, sequencing and DNA extraction (38). Of note, the costs varied quite significantly across sequencing platforms (38). Some of these costs, such as library prep, were greater for Illumina MiSeq ($28.56) than Nanopore MinION ($16.11) (Table 7C). On the other hand, sequencing had the greatest costs for Nanopore MinION ($47.08) when compared to Illumina MiSeq ($19.52) (Table 7C). Depth and breadth of coverage were identified as influential cost-driving with higher depth and breadths of coverage leading to higher costs (40).

3.1.1.4. Component Costs

Not including the BIA, four of the tNGS manuscripts investigated individual component costs (Tables 7B). These components included equipment, as well as reagents and consumables (Tables 7B). None of the manuscripts accounted for costs associated with property and lab or overhead. For equipment, it was shown that there was quite a significant range in costs for the different sequencing instruments, with the lowest cost being $1,000 for Nanopore MinION, followed by $125,000 for Illumina MiSeq, and the greatest cost being Illumina NextSeq at $210,000 - $275,000 (41). As mentioned earlier, these values are based on limited data, particularly for costing Nanopore

MinION. The cost of reagents and consumables also varied by identity of sequencing manufacturer, with the lowest being Illumina MiSeq100, and the highest Nanopore MinION. However, this was based on limited data (n=1). The per sample cost for reagents and consumables was found to be as low as $6.20 for MiSeq and $15.32 for iSeq100 (Table 7B). For this determination, samples had 15,000-fold coverage, which account for 99% of the genome, and a 30x-fold depth of coverage. For Chan *et al.* (2020), with Nanopore MinION, reagents and consumables were as high as $67.02 per sample per 24-plex workflow (Table 7B) (42). This included the cost of flow cell, barcoding kit, ligation sequencing kit, and other reagents (Table 7B). For Gliddon *et al.* (2021), reagents and consumables were as high as $138.68 per sample with Nanopore MinION (Table 7B) (43). This number was determined from samples with 1,000-10,000-fold coverage, sequences at 12 samples per run.

3.1.2. Budget Impact Analysis for tNGS and WGS

3.1.2.1. Summary of Findings

The manuscript by Cates *et al.* (2022) included a budget impact analysis (BIA) for Moldova, assessing several hypothetical scenarios for implementing tNGS (39). This included 7 tNGS, 4 WGS, as well as 1 pDST scenarios (39). Each scenario was measured across a 5-year timespan, and included total diagnostic but not treatment costs. The authors projected the expected number of samples for the years 2021-2025, and estimated the costs of introducing as well as routine use of tNGS or WGS (39). All costs were reported in 2021 USD. This was the only study that included a BIA.

3.1.2.2. Diagnostic Cost per Sample and Component Costs

Overall, scenarios with the lowest costs were those where tNGS or WGS were done in lieu of other testing, as opposed to scenarios where they were done in addition to other testing (Tables 8A-D). The authors found that the cost of tNGS per sample could be as low as $85.28 when looking at all positive cultures, and as high as $185.04 for positive cultures that had positive pDST and negative GeneXpert (Tables 8A-D). The manuscript by Cates L *et al.* (2022) also looked at the cost of labor, equipment, reagents & consumables, training, and maintenance as a percentage of total costs across all scenarios (39). For every tNGS, WGS and pDST scenario, reagents & consumables were the most significant cost category. Reagents and consumables costs ranged from $68.74-$128.97 for tNGS; $82.55-$135.16 for WGS; and $23.41 for pDST (Tables 8A-D).

3.1.2.3. Budget Impact Analysis

The authors projected cost across all five years was $361,695.00 for the pDST scenario (Tables 9A-D). Overall, the WGS scenarios were the most expensive, followed by the tNGS ones, and lastly pDST only. For the three WGS scenarios, the total cost for all five years was the lowest in the scenario where WGS followed positive cultures, excluding negative GeneXpert ($813,389.00), and greatest in the scenario where WGS was used with all positive cultures plus pDST ($1,485,961.00) (Tables 9A-D). For the 7 tNGS scenarios, the total cost for all five years was the lowest in the scenario where tNGS was performed on samples with a positive sputum smear microscopy (SSM) or GeneXpert but excluding rifampin susceptible cases as diagnosed on GeneXpert ($498,000.00). Total cost was greatest in the scenario where tNGS was used with all positive cultures plus pDST ($1,135,223.00) (Tables 9A-D). Of note, all of the tNGS (year 1 costs ranged from $136,029.00 to $271,341.00 and the years 2-5 costs ranged from $87,361.00 to $224,970.00), WGS (year 1 costs ranged from $297,132.00 to $421,963.00 and the years 2-5 costs ranged from $129,507.00 to $276,250.00) and pDST (year 1 costs were $76,811.00 and years 2-5 ranged from $68,001.00 to

$74,507.00) scenarios had additional projected start-up costs in year 1 due to capital expenditures and equipment costs (39).

The results of the economic evidence of diagnosing DR-TB with WGS are summarized in Appendix 8.1.

3.1.3. Quality of Health Economic Studies

Our quality of health economic studies extraction form was used to evaluate the quality of included tNGS (Supplementary Tables 1A and 1B) and WGS (Supplementary Tables 2A and 2B) manuscripts. This contained 17 questions, and was adapted from a modified CHEERS, as described earlier. Quality scores varied from 9/17 to 14/17 with an average quality score of 12/17 among tNGS papers. For WGS paper, quality scores varied from 7/17 to 15/17 with an average score of 11/17. The summary of this assessment is displayed graphically on Figure 3.

3.2 Scoping Review

3.2.1. Summary of Findings

Overall, there were 13 studies included in our scoping review (Figure 2) (47-59). Articles were excluded if they did not include persons with a disease of interest, with the exception of the two aforementioned papers by Alleweldt *et al.* (2021) and Dimitrov *et al.* (2017) (47 and 48). These two manuscripts assessed different disease entities but contained helpful data, and were thus included in our review. Further exclusion criteria were not using either tNGS or WGS as a diagnostic intervention; or not including any economic or cost data.

The 13 included studies, represented a variety of infectious pathogens, including avian influenza (n=1), influenza A&B (n=1), or food-borne pathogens (n=1); multiple avian paramyxoviruses (n=1); HIV (n=7); malaria (n=2); and HCV (n=3) (Table 3A-B). There were no included studies that considered HBV. There was a wide range of geographic locations among included studies: Europe and the Americas (n=1), India (n=1), USA (n=3), Brazil (n=1), Sweden (n=1), Kenya (n=1), Thailand (n=1), France (n=1), UK (n=1), Australia (n=1), and Tanzania (n=1) (Tables 3A and B). For diagnostic modalities, 9 of the studies looked at tNGS, and 4 looked at WGS (Tables 3A-B).

In terms of economic outcomes, 1 study looked at cost data and a breakeven analysis, one at micro and gross costing, 10 at some costing data, and 1 included a cost-effectiveness analysis (Tables 3A-B). Outcomes included component costs per sample (n=6), costs per step (n=3), ICERs (n=1), and cost per sample (n=4) (Tables 3A-B). The manuscript by Alleweldt *et al.* (2021) included a systematic review, and as a result, has a dedicated section in our results (47).

3.2.2. Costs per sample

In total, 4 of the 13 manuscripts assessed total cost per sample. There was a large range reported, from $24.33 per sample for Dudley *et al.* (2012) to $280.12 per sample for Gachogo *et al.* (2020) (50, 54). The paper by Dudley *et al.* (2012) only considered the cost of sequencing per sample, whereas the paper by Gachogo *et al.* (2020) performed a bottom-up costing that included broader operating costs (50, 54.

3.2.3. Component Costs

Excluding the paper by Alleweldt *et al.* (2021), of the 5 studies that reported component costs, all included the cost of reagents and consumables. For tNGS, this value ranged from $54.37 to

$104.61, both for HIV. The remaining studies were between $33.49 to $62.72 for reagent and consumables. Only the manuscript by Gachogo *et al.* (2020) assessed component costs per sample for other components in addition to reagents and consumables (50). These ranged from $2.44 for maintenance per sample, to $105.56 for equipment per sample. The paper by Merel *et al.* (2001) was the only one that assessed total capital costs per sequencer, but this was not incorporated into the per sample cost (52). These ranged in cost from $127,132.98 for the capillary electrophoresis (CEQ 2000) in-house protocol sequencer to $165,181.43 for the Prism-377 applied biosystems in-house protocol gel-based sequencer (52).

3.2.4. Costs per step

Three of the included studies considered costs per step. All three studies reported sequencing as the most expensive step. This ranged from $56.40 per sample (Dimitrov *et al.*) to $1093.49 per sample (Gachogo *et al*) (48, 50). The manuscript by Dimitrov *et al.* looked at multiple avian paramyxoviruses, and did not include all costs, such as labor costs (48). Meanwhile, the manuscript by Gachogo *et al*. (2020) looked at HIV and relied on a bottom-up costing analysis, which was more comprehensive (50). Additionally, the manuscript by Gachogo *et al.* (2020) was the most detailed and considered cost data for the following steps: sample collection, target enrichment multiplex PCR, sequencing, DNA and RNA extraction, gel electrophoresis, as well as sequencing analysis (50). In the Gachogo *et al.* (2020), sample collection had the lowest costs at $2.48 per sample, whereas sequencing had the highest costs at $165.88 per sample (50).

3.2.5. Incremental cost-effectiveness ratios (ICERs)

The paper by Weinstein *et al.* (2001) was the only one that performed a cost-effectiveness analysis, with ICERs as the main outcome (49). The authors compared the cost-effectiveness of genotypic resistance testing and clinical judgement versus clinical judgement alone for guiding antiretroviral therapy in patients with drug-resistant HIV. Overall, the QALYs in months ranged from 63.1-66.4 (49). The cost-effectiveness ratios were reported as $27,086 - $29,746 per QALY gained (49).

3.2.6. Output from Alleweldt et al. (2021)

The paper by Alleweldt *et al.* (2021) considered the costs and benefits of WGS through case studies across 8 different sites in Europe and the Americas (47). This involved looking at avian influenza (n=2), human influenza (n=1), and food-borne pathogens (n=5). The authors considered costs per sample and individual component costs. As a result of the amount of data available, we elected to present the findings of this paper in its own section (Tables 12A-E).

Overall, the costs per sample had a large range, from $62.85 for food-borne pathogens (INEI-ANLIS) to $1320.60 per sample for avian influenza (APHA). Component costs that were assessed included equipment, reagents and consumables, staffing, and other. The greatest costs were associated with reagents and consumables, which ranged from as low as $42.56 per sample (food-borne ARG) to as high as $1079.43 per sample (avian influenza APHA). The lowest costs were associated with other costs, and range from as low as $0.00 for every study with the exception of influenza A&B (IZSLER), which was $4.72 per sample. Factors identified by the authors to account for this substantial cost difference included the operational efficiency of the lab, the availability of trained personnel, sequencing technology used, the extent to which sequencers were used at full capacity, and discounts for capital costs as well as reagents and consumables when purchasing high enough volumes (47).

3.2.7. Quality of health economics studies

Our quality of health economic studies extraction form was used to evaluate the quality of included manuscripts (Supplementary Tables 3A and 3B). This contained 21 questions, and was adapted

from a modified CHEERS, as described earlier. Quality scores varied from 9/21 to 19/21 with an average quality score of 12/21 among. The summary of this assessment is displayed graphically on Figure 4.

4. DISCUSSION

4.1. Interpretation of Systematic Review Findings

To the best of our knowledge, this systematic review is the first of its kind to look at the economic evidence around using tNGS for the diagnosis of DR-TB. tNGS is a relatively new technology, and despite broad inclusion criteria, only ten suitable articles were found. There were no manuscripts with direct comparisons using a cost-effectiveness analysis, and included studies did not assess full end-to-end solutions, leading to a paucity of data.

The diagnostic cost per sample ranged between $69.64 to $73.47 for tNGS in the manuscript by Tafess K *et al.* (2020) (38). The lower value of this range was performed on Illumina MiSeq with 15,000-fold coverage and a batch size of 24 samples, whereas the upper end of this range was performed on Nanopore MinION with 15,000-fold coverage and a batch size of 12 samples. Both of these estimates included some equipment, as well as reagent and consumable costs, but did not account for labor or overhead costs. Both also used top-down costing. For WGS, this value ranged from $63.00 in Rowlinson MC and Musser KA (2022) to $277.00 in Vogel M *et al.* (2021) (41, 46). The lower range value was performed on Nanopore MinION; did not include certain costs like DNA extraction, preparation, or labor costs; and the authors used cross-batching. The upper number in this range was performed on Illumina MiSeq; incurred additional costs unique to the first 174 samples, such as errors or training inefficiencies; included many more costs, such as training, labor, and maintenance; and did not undergo cross-batching.

Sequencing was found to be the most costly step in both the tNGS or WGS approaches assessed, at approximately 28% of total costs for tNGS (n=1) and 46-56% of total costs for WGS (n=1). While component costs were not consistently reported, the greatest component cost was consistently reagents and consumables (n=5).

Five key cost drivers were identified that impact unit test costs associated with tNGS and WGS. These were inclusion of component costs in the cost per sample determination; using different sequencers; depth and breadth of coverage; inefficiencies in initial sample runs; and economics of scale via batching or cross-batching. Including more component costs in the cost per sample determination was shown to increase cost estimates. In terms of sequencers, using Nanopore MinION was associated with greater diagnostic costs per sample for tNGS, and using Illumina MiSeq was associated with greater diagnostic costs per sample for WGS. It should be noted that these were imperfect comparators, as it did not account for other differences in the papers, such as component costs included or samples per run. Sequencers are generally reported as one of the most expensive pieces of equipment needed for end-to-end sequencing, and as a result, is one of the only discrete pieces of equipment with costs included in these papers. For example, the manuscript by Rowlinson MC and Musser KA (2022) demonstrated that the Illumina MiSeq instrument costs $125,000, the Illumina NextSeq Instrument costs $210,000-275,000, and the Nanopore MinION instrument costs $1,000 (41). Of note, the Nanopore MinION has lower capital costs for the sequencer, but these are carried onto greater reagent and consumables costs. Third, greater depth and breadth of coverage is associated with greater costs, as this requires more genetic material to be sequenced. Fourth, inefficiencies unique to initial sample runs have been described to increase costs. Indeed, the manuscript by Vogel *et al.* (2021), demonstrated that the initial 174 runs were more costly than the latter runs, as skewed by “training inefficiencies” (46). This was the only tNGS or WGS manuscript that factored this into their cost determinations.

Lastly, lower sample volumes and less batching per run are likely to be associated with a greater costs per sample. This includes batching across TB samples or multiplexing with different organisms. While not explicitly mentioned in any of the papers, this could likely impact costs by spreading out certain overhead (e.g., labor costs) or equipment costs (e.g., sequencer flow cell) among more samples. This latter impact was discussed with the authors of one of the included manuscripts (41).

4.2. Interpretation of Scoping Review Findings

Similar to our systematic review and to the best of our knowledge, this is the first scoping review of its kind. There was likewise a relative paucity of data available with a total of 13 included studies. However, this did include a cost effectiveness analysis, (49).

Consistent with findings from the systematic review, the scoping review found the sequencing equipment was the only capital cost recorded, further supporting that it is likely the most expensive piece of equipment (Table 11C). Similarly, the greatest component costs were reagents and consumables, across included studies (n=6) (Tables 11A and 12B) with sequencing reported as the step with the greatest cost across included studies (n=3) (Table 11B). Lastly, the authors had identified similar cost drivers to the systematic review findings, including: economies of scale via batching or cross-batching/multiplexing, specific sequencing technology used, and including more cost components in unit test estimates.

The scoping review revealed additional cost factors that may impact cost and cost-effectiveness of screening including: operational efficiency of the lab, availability of trained personnel, sequencers being used at full capacity, discounts associated with purchasing high volume from the same supplier, as well as complexity of infectious pathogen.

4.3. Limitations of the Reviews

There are several limitations worth noting with the systematic review, many resulting from the limited data available. First, most of the studies were based in North America (n=5) or Western Europe (n=3) with fewer in LMICs (n=4), making the findings less generalizable to these regions. For the tNGS only, both tNGS and WGS, and WGS only manuscripts; 66%, 66%, and 50% of manuscripts were in the US or Europe, respectively (Tables 1 and 2A-B). Second, there was significant heterogeneity between study methods, and it was not always clear whether diagnostic cost per sample across studies included the same component costs. Third, cost reporting was limited to cost per sample, cost per step, component costs, as well as a single BIA; and no direct cost-effectiveness analyses were found. Fourth, there was quite a range in study quality for both tNGS (9-14/17) and WGS (7-15/17) papers (Figure 3). However, it is worth mentioning that the original CHEERS criteria are meant for cost effectiveness studies, for which there were none included in the systematic review. For the most part, similar limitations were identified with the scoping review.

4.4. Additional Study

An additional study by Mugwagwa *et al.* (2021) was identified that was not captured in our search strategy as it targeted DS-TB and therefore did not meet our study inclusion criteria (60). Here, the authors used an integrated transmission-dynamic health economic model to perform a cost-effectiveness analysis of molecular testing and/or WGS for the diagnosis of TB in a low-burden setting. Overall, the authors found that the cost per TB case detected with either molecular testing or WGS was cost-effective when compared to culture-based testing or using chest x-ray. Furthermore, the combined use of molecular testing and WGS was the most cost-effective strategy. This was due to the greater sensitivity of molecular testing and WGS, as well as shortened time to diagnosis. The latter saved costs tied to reduced transmission of TB, shorter morbidity, reduced future treatment costs, and fewer hospitalizations.

5. CONCLUSION

Given the growing interest in tNGS as a diagnostic modality, there are a number of additional data points that would be helpful to characterize. First, further studies are needed to determine additional economic markers, such as cost-effectiveness analyses or ICERs. Second, it will be prudent to develop a better understanding of the various factors that may impact cost, such as the impact of batching and multiplexing, as well as country-dependent factors. Lastly, it would be helpful to have more data points that include all relevant costs, such as overhead, total lab equipment, and staffing.

In conclusion, our systematic and scoping reviews are the first of their kind to assess the economic evidence around tNGS and WGS for the diagnosis of DR-TB. All of the studies from the systematic review ere quite recent (2018-2021), reflecting that this is a developing field for tuberculosis diagnostics. Further studies are needed to determine additional economic markers, such as cost-effectiveness analyses or ICERs; how various factors may impact cost, such as depth of coverage; as well as including all relevant costs, such as overhead and total lab equipment in cost calculations.

6. TABLES & FIGURES

Figure 1. PRISMA Flow Diagram for the Systematic Review

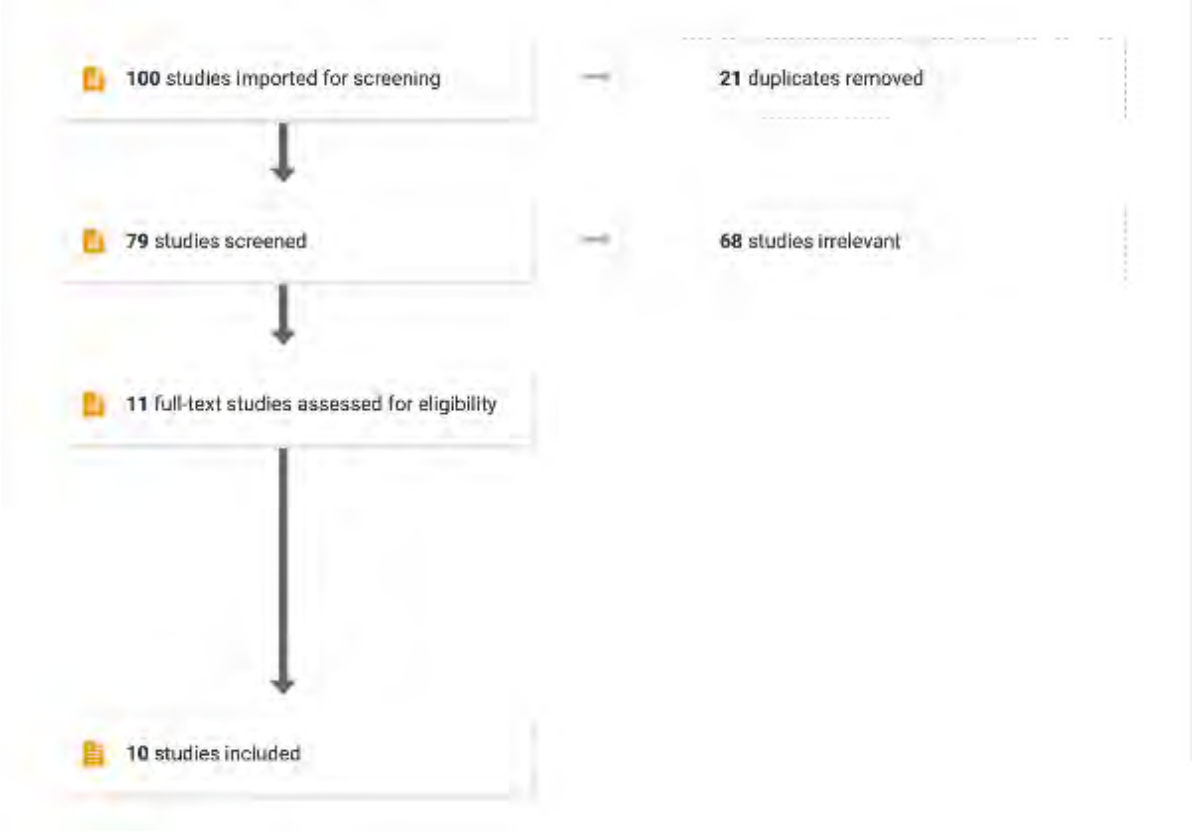


Figure 2. PRISMA Flow Diagram for the Scoping Review

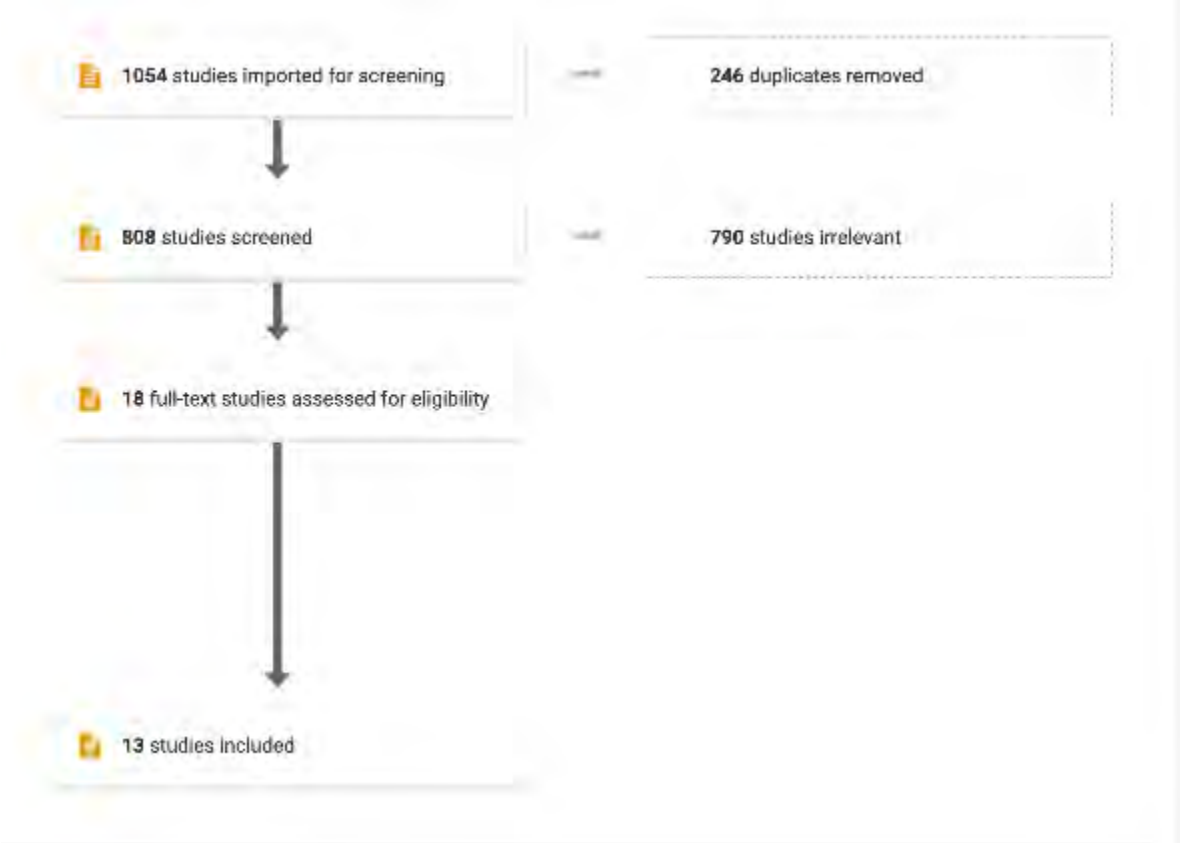


Table 1. Characteristics of Included tNGS Studies for the Systematic Review

Manuscript	Country setting	Year of Cost Valuation	Currency	Study Population	Diagnostic Strategies (tNGS or WGS)	Reference Diagnostic Strategies (pDST or mWRDs)	Analysis Perspective (Healthcare or Societal)	Type of Economic Evaluation	Primary Outcome	Secondary Outcome(s)	Source of Costing
Colman RE *et al* (2019)	USA	2018	USD	TB isolates	tNGS (iSeq100 and miSeq) and WGS (iSeq100 and MiSeq)	None	Healthcare	Some cost data	Cost per sample	Capital costs of sequencing instruments	US list pricing as of October, 2018
Chan WS *et al* (2020)	USA	2020	USD	General population with AFB smear-negative or MTB detected low/very low	tNGS (MinION and MiSeq)	pDST	Healthcare	Some cost data	Reagent cost per sample	None	Adapted from Oxford Nanopore Technologies company website
Tafess K *et al* (2020)	Hong Kong and Ethiopia	2019	HKD	TB isolates	tNGS (in-house MiSeq and MinION)	pDST	Healthcare	Some cost data	Cost per sample	Cost per stage of tNGS	Hong Kong list prices as of 2019
Gliddon HD *et al* (2021)	UK/South Africa	2020	GBP	TB isolates	tNGS (MinION) and WGS (MiSeq)	pDST	Healthcare	Some cost data	Consumables cost per sample	None	Adapted From Votintseva et al (2016) manuscript
Cates L *et al* (2022)	Moldova	2021	USD	General population	tNGS (iSeq100) and WGS (MiSeq)	pDST and mWRDs	Healthcare	Budget impact analysis (BIA)	Cost per 5-year period for population	Cost per sample	Published sources, manufacturer costing data, publicly available costing data
Rowlinson MC and Musser KA (2022)[1]	USA	Not stated	USD	Not stated	tNGS and WGS (both in-house)	pDST and mWRDs	Healthcare	Some cost data	Cost per sample	Capital costs of sequencing instruments	In-house laboratory costs

1. The paper by Rowlinson MC and Musser KA (2020) is a report

Table 2A. Characteristics of Included WGS Studies for the Systematic Review

Manuscript	Country Setting	Year of Cost Valuation	Currency	Study Population	Diagnostic Strategies (tNGS or WGS)	Reference Diagnostic Strategies (pDST or mWRDs)
Cirillo DM *et al* (2016)	Italy	2016	EUR	Confirmed MDR/RR-TB cases in general population	WGS	Standard testing
Pankhurst LJ *et al* (2016)	Europe (UK, Ireland, Germany and France) and North America (Canada)	2014	GBP	Confirmed MDR/RR-TB cases in general population	WGS	Routine MTBC diagnostic workflows
Colman RE *et al* (2019)	USA	2018	USD	TB isolates	tNGS (in-house iSeq100 and MiSeq) and WGS (in-house iSeq100 and MiSeq)	None
He G *et al* (2020)	China	2016	CNY	Admitted patients age >18 years of age, confirmed MDR-TB	WGS	pDST
Vogel M *et al* (2021)	Kyrgyz Republic	2021	USD	Confirmed MDR/RR-TB cases in general population	WGS (MiSeq)	pDST
Rowlinson MC and Musser KA (2022)[1]	USA	Not stated	USD	Not stated	tNGS and WGS	pDST and mWRDs

1. The paper by Rowlinson MC and Musser KA (2020) is a report

Table 2B. Characteristics of Included WGS Studies for the Systematic Review

Manuscript	Analysis Perspective (Healthcare or Societal)	Type of Economic Evaluation	Primary Outcome	Secondary Outcome(s)	Source of Costing
Cirillo DM *et al* (2016)	Healthcare	Some cost data	Cost per sample	N/A	N/A
Pankhurst LJ *et al* (2016)	Healthcare	Bottom-up costing via micro-costing	Cost per sample	N/A	Questionnaires, expert consultations, and interviews with laboratory staff
Colman RE *et al* (2019)	Healthcare	Some cost data	Cost per sample	Capital costs of sequencing instruments	United States list pricing as of October, 2018
He G *et al* (2020)	Healthcare	Some cost data	Cost per sample	N/A	N/A
Vogel M *et al* (2021)	Healthcare	Some cost data	Cost per sample	Component costs	N/A
Rowlinson MC and Musser KA (2022)[1]	Healthcare	Some cost data	Cost per sample	Capital costs of sequencing instruments	In-house laboratory costs

1. The paper by Rowlinson MC and Musser KA (2020) is a report

Table 3A. Characteristics of Included Studies for the Scoping Review

Manuscript	Communicable disease	Country setting	Laboratory classification (peripheral, intermediate or central)	Year of cost valuation	Currency	Study population	Diagnostic strategies (tNGS or WGS)	Reference diagnostic strategies	Analysis perspective (healthcare or societal)	Type of economic evaluation	Type of model if applicable	WTP threshold if applicable	Primary outcome	Secondary outcome(s)	Source of costing
Alleweldt et al (2021)	Avian influenza; influenza A&B; food-borne pathogens	Europe and the Americas	Central	2016-2019	Euro	Clinical isolates	WGS	Conventional Methods	Healthcare	Cost data and breakeven analysis	N/A	N/A	Breakeven analysis	Cost per sample and component costs	Original purchase cost of equipment and consumables; country-specific labor costs
Chaturbhuj et al (2014)	HIV	India	Peripheral	Not stated	USD	Panel sequences	tNGS (in-house)	tNGS (ViroSeq Genotyping System 2.0)	Healthcare	Some cost data	N/A	N/A	Cost per test	N/A	N/A
Dimitrov et al (2017)	Multiple avian paramyxoviruses	USA	Central	Not stated	USD	Clinical isolates	tNGS (Illumina MiSeq)	N/A	Healthcare	Some cost data	N/A	N/A	Cost per sample	Cost per steps	N/A
Dudley et al (2012)	HIV	Brazil	Central	Not stated	USD	General population	tNGS (Roche/454)	tNGS (Sanger sequencing)	Healthcare	Some cost data	N/A	N/A	Cost per sample	N/A	N/A
Ekici et al (2014)	HIV	Sweden	Central	2014	USD	Therapy-naïve HIV1-infected patients	tNGS (Illumina MiSeq)	tNGS (Standard population sequencing)	Healthcare	Some cost date	N/A	N/A	Cost per sample	N/A	N/A
Gachogo et al (2020)	HIV	Kenya	Central	2019	USD	Not stated	tNGS (HIVDR)	tNGS (FDA-approved ViroSeq HIV genotyping assay)	Healthcare	Micro and gross costing	N/A	N/A	Cost per sample	Component costs and costs per steps	Interviews with laboratory and management staff; quotations, invoices and delivery notes
Kunasol et al (2022)	Malaria	Thailand	Central	Not stated	USD	Patients with uncomplicated falciparum malaria	tNGS (Ion Torrent PGM)	tNGS (Illumina MiSeq)	Healthcare	Some cost data	N/A	N/A	Total cost of consumables	N/A	N/A

Table 3B. Characteristics of Included Studies for the Scoping Review

Manuscript	Communicable disease	Country setting	Laboratory classification (peripheral, intermediate or central)	Year of cost valuation	Currency	Study population	Diagnostic strategies (tNGS or WGS)	Reference diagnostic strategies	Analysis perspective (health care or societal)	Type of economic evaluation	Type of model if applicable	WTP threshold if applicable	Primary outcome	Secondary outcome(s)	Source of costing
Merel et al (2001)	HIV	France	Central	Not stated	USD	HAART-treated patients with virological failure	tNGS with Capillary electrophoresis (CEQ 2000 sequencer)	tNGS with Gel plate-based sequencing (Prism-377 and TruGene kit)	Healthcare	Some cost data	N/A	N/A	Reagent costs per sequence	Sequencer costs	French price listing
Patel et al (2016)	HIV and HCV	UK	Not stated	Not stated	GBP	Samples from UK population	WGS	tNGS (Sanger sequencing)	Healthcare	Some cost data	N/A	N/A	Mean cost per sample	N/A	Published genomic-testing costing templates
Riaz et al (2021)	HCV	Australia	Central	2019	AUD	Clinical isolates	WGS (Oxford Nanopore)	WGS (Illumina)	Healthcare	Some cost data	N/A	N/A	Reagent costs per sequence	N/A	Australian Reagent costs as of 2019
Taylor et al (2013)	Malaria	Tanzania	Central	Not stated	USD	Children with uncomplicated *P. falciparum* malaria	tNGS (Second Generation Sequencing)	tNGS (Sanger sequencing)	Healthcare	Some cost data	N/A	N/A	Cost per specimen	Cost per step	N/A
Wales et al (2017)	HCV	USA	Central	2016	USD	Plasma samples	WGS	WGS (Sanger sequencing)	Healthcare	Some cost data	N/A	N/A	Cost of reagents per sample	N/A	List prices
Weinstein et al (2001)	HIV	USA	Not stated	1998	USD	HIV-infected patients in US with baseline CD4 counts of 0.250x10^9 cells/L.	tNGS (Genotypic resistance testing) and clinical judgement	Clinical judgement	Societal	Cost effectiveness analysis	State transition model, first order Monte-Carlo simulation	N/A	Cost-effectiveness per QALY gained	Life expectancy, quality-adjusted life expectancy	AIDS Cost and Services Utilization Survey (ACSUS) public use data tapes; the 1998 Red Book

Table 4. Methodologies of Included tNGS Studies for the Systematic Review

Manuscript	Depth for costing	Coverage for costing	Varying Sample Numbers for Costing	Different Scenarios Modelled for Costing	Key Scenarios/Variables Explored in Sensitivity Analyses (PSA, One Way, or Two Way)	Time to Completion
Colman RE et al (2019)	Yes	No	Yes	No	No	MiSeq 24 hours; iSeq100 17.5 hours for sequencing
Chan WS et al (2020)	No	No	No	No	No	MiSeq 48.05 hours; MinION 8.68 hours; pDST 69.5 days for end-to-end solution
Tafess K et al (2020)	No	No	No	No	No	MiSeq 38 hours; MinION 15 hours; pDST 13-21 days for end-to-end solution
Gliddon HD et al (2021)	No	No	No	No	No	None
Cates L et al (2022)	No	No	Yes	Yes	Yes (one-way sensitivity analysis)	None
Rowlinson MC and Musser KA (2022)[1]	No	No	No	No	No	MiSeq 40 hours; NextSeq 24 hours; MinION 24-48 hours for sequencing

1. The paper by Rowlinson MC and Musser KA (2020) is a report

Table 5. Methodologies of Included WGS Studies for the Systematic Review

Manuscript	Depth for costing	Coverage for costing	Varying sample numbers for costing	Different scenarios modelled for costing	Key scenarios/variables explored in sensitivity analyses (PSA, one way, or two way)	Time to completion
Cirillo DM et al (2016)	No	No	No	No	No	72 hours
Pankhurst LJ et al (2016)	No	No	Yes	Yes	No	216 hours
Colman RE et al (2019)	Yes	No	Yes	No	No	MiSeq 24 hours; iSeq100 17.5 hours for sequencing only
He G et al (2020)	No	No	No	No	No	168 hours
Vogel M et al (2021)	No	No	Yes	No	No	N/A
Rowlinson MC and Musser KA (2022)[1]	No	No	No	No	No	MiSeq 40 hours; NextSeq 24 hours; MinnION 24-48 hours for sequencing only

1. The paper by Rowlinson MC and Musser KA (2020) is a report

Table 6. Methodologies of Included Studies for the Scoping Review

Manuscript	Depth for costing	Coverage for costing	Varying sample numbers for costing	Different scenarios modelled for costing	Key scenarios/variables explored in sensitivity analyses (PSA, one way, or two way)	Did authors include information on feedback and/or knowledge translation?	Time to completion
Alleweldt et al (2021)	No	No	No	No	No	No	N/A
Chaturbhuj et al (2014)	No	No	No	No	No	No	N/A
Dimitrov et al (2017)	No	No	No	No	No	No	25-30 hours
Dudley et al (2012)	No	No	No	No	No	No	N/A
Ekici et al (2014)	No	No	No	No	No	No	N/A
Gachogo et al (2020)	No	No	No	No	Yes (one-way sensitivity analysis for 20% variations to cost categories)	No	N/A
Kunasol et al (2022)	No	No	No	No	No	No	N/A
Merel et al (2001)	No	No	No	No	No	No	108-189 hours[1]
Patel et al (2016)	No	No	No	No	No	No	N/A
Riaz et al (2021)	No	No	No	No	No	No	47 hours
Taylor et al (2013)	No	No	Yes	No	No	No	N/A
Wales et al (2017)	No	No`	Yes	No	No	No	6-14 hours[2]
Weinstein et al (2001)	No	No	No	Yes	Yes	No	N/A

1. Incudes sequencing and data analysis steps, range is 4.5 days for CEQ 2000 and 8 days for OpenGene
2. Only includes time needed for sequencing, with 363 minutes for 30 samples, and 826 minutes for 96 samples

Table 7A. Output from Included tNGS Studies for the Systematic Review: Diagnostic Cost per Sample

Manuscript	Diagnostic Cost per Sample[1,2,3]	
	Diagnostic Intervention (tNGS)	
	Illumina MiSeq (15,000x coverage)	Nanopore MinION
Tafess K *et al* (2020)	$69.64	$73.47

1. Values are written in USD 2021 unless otherwise specified
2. Value looks at Illumina MiSeq and for lower and upper values, respectively, both at 15,000x coverage
3. Only studies with published data for each outcome were included

Table 7B. Output from Included tNGS Studies for the Systematic Review: Component Costs

Manuscript	Component costs[1, 4]					
	Diagnostic Intervention (tNGS)					
	Illumina MiSeq Instrument	Illumina NextSeq Instrument	ONT MinION Instrument	Reagents and consumables	Reagents (Nanopore MinION, 24-plex workflow)	Consumables
Colman RE *et al* (2019)				$6.20-15.32[2]		
Chan WS *et al* (2020)					$67.02	
Gliddon HD *et al* (2021)						$138.68[2]
Rowlinson MC and Musser KA (2022)[3]	$125,000	$210,000 - $275,000	$1,000			

1. Values are written in 2021 USD unless otherwise specified
2. Value includes: £9.70 Moyi DNA extraction; £2.50 RPA reagents; £5.20 FFPE DNA repair mix & buffer; £7.20 Ultra II End prep mix & buffer; £7.60 blunt/TA ligase master mix; £2.15 quick ligation module; £6.66 ligation sequencing kit; £3.26 native barcoding kit; £52.92 flow cell; and £1.20 agencourt beads
3. Cost was calculated as a rough estimation, which includes flow cell ($900.00), barcoding kit ($1200.00), ligation sequencing kit ($599.00), and other reagents. It does not include salary and other fixed costs
4. Only studies with published data were for each outcome were included

Table 7C. Output from Included tNGS Studies for the Systematic Review: Costs per step

Manuscript	Costs per step[1,3]									
	Diagnostic Intervention (tNGS Illumina MiSeq)					*Diagnostic Intervention (Nanopore MinION)*				
	Target Enrichment Multiplex PCR	Library Prep	Quality Control	Sequencing	DNA Extraction	Target Enrichment Multiplex PCR	Library Prep	Quality Control	Sequencing	DNA Extraction
Tafess K *et al* (2020)[2]	$4.15	$28.56	$13.78	$19.52	$3.64	$4.15	$16.11	$2.52	$47.08	$3.64

1. Values are written in 2021 USD unless otherwise specified
2. Value looks at Illumina MiSeq and Nanopore MinION for lower and upper values, respectively, both at 15,000x coverage
5. Only studies with published data for each outcome were included

Table 8A. Output from Included tNGS and WGS Studies for the Systematic Review: Costing Data, Cates et al. (2022)

Manuscript	Diagnostic cost per sample[1,2]																	
	Diagnostic Intervention (tNGS Illumina iSeq100)																	
	SSM+ or Xpert+, excluding Xpert/RIF-						All SSM+ or Xpert+						All SSM+ or Xpert+, plus pDST					
	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total
Cates L et al (2022)	$1.40	$19.34	$97.94	$0.70	$20.74	$140.12	$0.64	$8.84	$71.93	$0.36	$9.48	$91.17	$7.10	$18.58	$99.19	$0.96	$10.93	$136.63

1. Values are written in 2021 USD unless otherwise specified
2. Component costs determined by multiplying proportions with cost per sample, verified by authors

Table 8B. Output from Included tNGS and WGS Studies for the Systematic Review: Costing Data, Cates et al. (2022)

Manuscript	Diagnostic cost per sample[1,2]																	
	Diagnostic Intervention (tNGS Illumina iSeq100)																	
	All culture+						All culture+ excluding Xpert/RIF-						All culture+, plust pDST					
	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total
Cates L et al (2022)	$0.51	$7.59	$68.74	$0.26	$8.10	$85.28	$1.04	$14.16	$85.17	$0.46	$15.20	$116.03	$7.01	$15.90	$92.24	$0.88	$9.26	$125.16

1. Values are written in 2021 USD unless otherwise specified
2. Component costs determined by multiplying proportions with cost per sample, verified by authors

Table 8C. Output from Included tNGS and WGS Studies for the Systematic Review: Costing Data, Cates et al. (2022)

Manuscript	Diagnostic cost per sample[1,2]																	
	Diagnostic Intervention (tNGS Illumina iSeq100)						Diagnostic Intervention (WGS Illumina MiSeq)											
	All culture+ excluding Xpert/RIF-, plus pDST						All culture+						All culture+ excluding Xpert/RIF-					
	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total
Cates L et al (2022)	$7.40	$29.61	$128.97	$1.48	$17.39	$185.04	$1.98	$26.15	$82.55	$0.25	$13.01	$123.95	$2.52	$48.97	$91.40	$0.50	$24.32	$167.70

1. Values are written in 2021 USD unless otherwise specified
2. Component costs determined by multiplying proportions with cost per sample, verified by authors

Table 8D. Output from Included tNGS and WGS Studies for the Systematic Review: Costing Data, Cates et al. (2022)

Manuscript	Diagnostic cost per sample[1,2]																	
													Diagnostic Comparitor (pDST)					
	All culture+, plust pDST						All culture+ excluding Xpert/RIF-, plus pDST						N/A					
	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total	Labor	Equipment	Reagents & Consumables	Training	Maintenance	Total
Cates L et al (2022)	$8.36	$34.57	$105.83	$0.82	$14.25	$163.83	$8.99	$64.39	$135.16	$1.66	$26.51	$236.71	$6.38	$8.30	$23.41	$0.56	$1.20	$39.88

1. Values are written in 2021 USD unless otherwise specified
2. Component costs determined by multiplying proportions with cost per sample, verified by authors

Table 9A. Output from Included tNGS and WGS Studies for the Systematic Review: BIA, Cates et al. (2022)

Manuscript	Total Budget Impact Estimates[1]																	
	pDST						tNGS for all culture+						tNGS for all culture+, excluding Xpert/RIF-					
	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total
Cates L *et al* (2022)	$76,811.00	$74,507.00	$72,272.00	$70,104.00	$68,001.00	$361,695.00	$194,530.00	$150,463.00	$146,577.00	$142,808.00	$139,151.00	$773,528.00	$149,772.00	$107,047.00	$104,464.00	$101,958.00	$99,528.00	$562,769.00

1. Values are written in 2021 USD unless otherwise specified

Table 9B. Output from Included tNGS and WGS Studies for the Systematic Review: BIA, Cates et al. (2022)

Manuscript	Total Budget Impact Estimates[1]																	
	tNGS for all SSM+ or Xpert+						tNGS for all SSM+ or Xpert+, excluding Xpert/RIF-						tNGS for all SSM+ or Xpert+, plus pDST					
	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total
Cates L *et al* (2022)	$180,787.00	$137,132.00	$133,646.00	$130,265.00	$126,985.00	$708,816.00	$136,029.00	$93,717.00	$91,534.00	$89,416.00	$87,361.00	$498,057.00	$255,837.00	$209,931.00	$204,261.00	$198,762.00	$193,427.00	$1,062,218.00

1. Values are written in 2021 USD unless otherwise specified

Table 9C. Output from Included tNGS and WGS Studies for the Systematic Review: BIA, Cates et al. (2022)

Manuscript	Total Budget Impact Estimates[1]																	
	tNGS for all culture+ plus pDST						tNGS for all culture+, excluding Xpert/RIF-, plus pDST						WGS for all culture+					
	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total
Cates L *et al* (2022)	$271,341.00	$224,970.00	$218,849.00	$212,911.00	$207,152.00	$1,135,223.00	$220,847.00	$175,991.00	$171,339.00	$166,827.00	$162,450.00	$897,456.00	$345,152.00	$201,743.00	$197,006.00	$192,411.00	$187,954.00	$1,124,267.00

1. Values are written in 2021 USD unless otherwise specified

Table 9D. Output from Included tNGS and WGS Studies for the Systematic Review: BIA, Cates et al. (2022)

Manuscript	Total Budget Impact Estimates[1]																	
	WGS for all culture+ excluding Xpert/RIF-						WGS for all culture+, plus pDST						WGS for all culture+, excluding Xpert/RIF-, plus pDST					
	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total	Year 1 (start-up)	Year 2	Year 3	Year 4	Year 5	Total
Cates L *et al* (2022)	$297,132.00	$137,704.00	$134,888.00	$132,157.00	$129,507.00	$813,389.00	$421,963.00	$276,250.00	$269,278.00	$262,515.00	$255,955.00	$1,485,961.00	$350,208.00	$206,648.00	$201,764.00	$197,026.00	$192,430.00	$1,148,075.00

1. Values are written in 2021 USD unless otherwise specified

Table 10A. Output from Included WGS Studies for the Systematic Review: Diagnostic Cost per Sample

Manuscript	Diagnostic cost per sample[1,7]			
	Diagnostic Intervention (WGS)			Diagnostic Comparator
	Illumina NextSeq	Illumina MiSeq	Nanopore MinION	pDST
Cirillo DM *et al* (2016)[10]		$185.77		
Pankhurst LJ *et al* (2016)[5]		$110.87 ($109.67-112.38)[4]		$49.00 ($48.60-49.47)[4]
He G *et al* (2020)		$71.61		$45.47-102.30[3]
Vogel M *et al* (2021)		$141.00 - $277.00[6]		
Rowlinson MC and Musser KA (2022)[2]	$68.00	$130.00 - $150.00	$63.00	

1. Values are written in 2021 USD unless otherwise specified.
2. Diagnostic costs per sample do not include DNA extraction, preparation costs, or labor costs.
3. First line drugs and second line drugs recorded as lower and upper ranges, respectively.
4. Range based on sensitivity analysis of 10% greater or fewer samples per year.
5. Component costs are calculated on a per sample basis.
6. Range based on first 174 sample ($277.00), and subsequent samples for 500-cycles ($167.00) and 600-cycles ($141.00).
7. Only studies with published data were for each outcome were included.

Table 10B. Output from Included WGS Studies for the Systematic Review: Component Costs

Manuscript	Component costs[1,4]													
	Diagnostic Intervention (WGS)									*Diagnostic Comparator (pDST)*				
	Equipment	MiSeq	iSeq100	Maintenance	Labor	Training	Overhead	Reagents and consumables (Illumina MiSeq)	Reagents and consumables (Illumina iSeq100)	Equipment	Labor	Training	Overhead	Reagents and consumables
Pankhurst LJ *et al* (2016)[3]	$10.48				$18.72	$0.68	$18.48	$62.51		$3.12	$17.61	$0.42	$8.17	$19.68
Colman RE *et al* (2019)		$106,820.31	$20,500.87					$15.08-$30.44[2]	$74.93					
Vogel M *et al* (2021)	$222,065.00			$8,462.00		$48,250.00		$60,995.66						

1. Values are written in 2021 USD unless otherwise specified.
2. Illumina MiSeq v3 2x300bp run - Illumina MiSeq v3 2x150bp run.
3. Component costs are calculated on a per sample basis.
4. Only studies with published data were for each outcome were included.

Table 10C. Output from Included WGS Studies for the Systematic Review: Costs per step

Manuscript	Costs per step[1,3]				
	Diagnostic Intervention (WGS)				
	DNA Extraction	Target Enrichment Multiplex PCR	Library Prep	Quality Control	Sequencing
Vogel M *et al* (2021)[2]	$4.05	$1.31 - $2.88	$59.61 - $91.25	$3.83 - $14.70	$64.53 - $156.32

1. Values are written in 2021 USD unless otherwise specified.
2. Range based on first 174 sample ($277.00), and subsequent samples for 500-cycles ($167.00) and 600-cycles ($141.00).
3. Only studies with published data were for each outcome were included.

Table 11A. Output from Included Studies for the Scoping Review: Key Outcome(s) and Component Costs

Manuscript	Key outcome(s)	Component costs per sample[1,5]					
		Diagnostic Intervention					
		Equipment	Maintenance	Staffing	Overhead	Reagents and consumables	Quality Assurance
Gachogo et al (2020)	To establish detailed cost profile for HIVDR testing and to identify cost drivers	$105.56	$2.44	$48.25	$15.14	$104.61	$4.12
Kunasol et al (2022)[2]	Targeted Amplicon Deep sequencing (TADs) using Ion Torrent PGM with Illumina MiSeq reduced costs by 86% compared to conventional Sanger sequencing					$62.72	
Merel et al (2001)	Reagent costs per sequencing reaction and sequencer capital costs were lower for CEQ 2000 than Prism-377 and TruGene					$54.37	
Riaz et al (2021)[3]	Reagent cost per sample was less for Nanopore sequencing than for Illumina sequencing					$33.49	
Wales et al (2017)	Reagent cost per sample					$19.19 - $36.13[4]	

1. All costs are in USD unless stated otherwise.
2. Costs are for 6 drug resistance genes per sample, with 96 samples per sequencing run.
3. Intervention was Nanopore and comparator was Illumina Sequencing.
4. The range is based on number of samples, where 30 samples were $36.13 per sample and 96 samples were $19.19 per sample.
5. Only manuscripts with data were included.

Table 11B. Output from Included Studies for the Scoping Review: Costs per step

Manuscript	Costs per step[1,2]					
	Diagnostic Intervention					
	Sample collection	Target Enrichment Multiplex PCR	Sequencing	DNA or RNA Extraction	Gel electrophoresis	Sequencing analysis
Dimitrov et al (2017)		$55.29	$56.40	$5.53		
Gachogo et al (2020)	$2.48	$57.86	$165.88	$23.49	$10.66	$19.75
Taylor et al (2013)		$1.99	$1093.49	$1.99		

1. All costs are in USD unless stated otherwise.
2. Only manuscripts with data were included.

Table 11C. Output from Included Studies for the Scoping Review: ICERs, Cost per Sample and Component Costs Total

Manuscript[1,7]	ICERs						Cost per Sample		Component costs total	
	Diagnostic Intervention						*Diagnostic intervention*	*Diagnostic Comparator*	*Diagnostic intervention*	*Diagnostic Comparator*
	QALYs (months)	QALYs sens analysis - effectiveness of HAART (months)	Added QALYs sens analysis - maximum duration of HAART (months)	Cost-effectiveness ratio (CER, $/QALY gained)	CER sens analysis - effectiveness of HAART ($/QALY gained)	CER sens analysis - maximum duration of HAART ($/QALY gained)			Equipment	Equipment
Chaturbhuj et al (2014)							$128.19	$343.37		
Dimitrov et al (2017)							$117.22			
Dudley et al (2012)[4]							$24.33	$69.35		
Ekici et al (2014)[5]							$27.18-$36.83			
Gachogo et al (2020)							$280.12	$391.10		
Merel et al (2001)[2]									$127,132.98	$165,181.43
Patel et al (2016)							$180.75	$119.99-$270.36[3]		
Weinstein et al (2001)[6]	63.1-66.4	60.9-66.3	2.20-2.69	$27,086 – 29,746	$26,754 – 39,882	$29,746 – 33,734				

1. All costs are in USD unless stated otherwise.
2. Equipment costs only include cost of sequencers; comparator only includes Prism-377 and not OpenGene.
3. Costs include HIV (79 GBP) and HCV (178 GBP) for Sanger sequencing, and 119 GBP for HIV/HCV for NGS.
4. Cost per sample assumes 48 samples multiplexed together.
5. Assumes pooling 24 samples per run.
6. Range presents data from two clinical trials.
7. Only manuscripts with data were included.

Table 12A. Output from Included Studies for the Scoping Review: Cost per Sample, Alleweldt et al. (2021)

Manuscript	Cost per sample															
	WGS								**Conventional**							
	Avian influenza (APHA)	Avian influenza (FLI)	Influenza A+B (EMC)	Food-borne Pathogens (IZSLER)	Food-borne Pathogens (INEI-ANLIS)	Food-borne Pathogens (MDH)	Food-borne Pathogens (PHAC)	Food-borne Pathogens (PHE)	Avian influenza (APHA)	Avian influenza (FLI)	Influenza A+B (EMC)	Food-borne Pathogens (IZSLER)	Food-borne Pathogens (INEI-ANLIS)	Food-borne Pathogens (MDH)	Food-borne Pathogens (PHAC)	Food-borne Pathogens (PHE)
Alleweldt *et al* (2021)[1]	$1320.60	$698.73	$126.40	$483.51	$62.85	$193.5	$273.26	$161.84	$376.81	$ 1027.72	$106.99	$112.42	$18.96	$101.46	$119.64	$85.03

1. Costs were obtained between 2016-2019 for all studies within Alleweldt et al (2021), with the majority of costs from 2017, which was used for costing conversions to 2021 USD.

Table 12B. Output from Included Studies for the Scoping Review: Component Costs, Alleweldt et al. (2021)

Manuscript	Component Costs															
	WGS								**Conventional Method**							
	Equipment: Avian Influenza APHA	Equipment: Avian Influenza FLI	Equipment: Influenza A+B EMC	Equipment: Food-borne pathogens IZSLER	Equipment: Food-borne pathogens ARG	Equipment: Food-borne pathogens MDH	Equipment: Food-borne pathogens PHAC	Equipment: Food-borne pathogens PHE	Equipment: Avian Influenza APHA	Equipment: Avian Influenza FLI	Equipment: influenza A+B EMC	Equipment: Food-borne pathogens IZSLER	Equipment: Food-borne pathogens ARG	Equipment: Food-borne pathogens MDH	Equipment: Food-borne pathogens PHAC	Equipment: Food-borne pathogens PHE
Alleweldt *et al* (2021)[1]	$76.03	$259.04	$3.21	$200.05	$17.50	$36.92	$96.31	$45.76	$102.04	$168.85	$3.41	$31.86	N/A	$7.30	$15.61	$9.24

1. Costs were obtained between 2016-2019 for all studies within Alleweldt et al (2021), with the majority of costs from 2017, which was used for costing conversions to 2021 USD.

Table 12C. Output from Included Studies for the Scoping Review: Cost per Sample, Alleweldt et al. (2021)

Manuscript	Component Costs															
	WGS								**Conventional Method**							
	Consumables: Avian influenza APHA	Consumables: Avian influenza FLI	Consumables: Influenza A+B EMC	Consumables: Food-borne pathogens IZSLER	Consumables: Food-borne pathogens ARG	Consumables: Food-borne pathogens MDH	Consumables: Food-borne pathogens PHAC	Consumables: Food-borne pathogens PHE	Consumables: Avian Influenza APHA	Consumables: Avian Influenza FLI	Consumables: influenza A+B EMC	Consumables: Food-borne pathogens IZSLER	Consumables: Food-borne pathogens ARG	Consumables: Food-borne pathogens MDH	Consumables: Food-borne pathogens PHAC	Consumables: Food-borne pathogens PHE
Alleweldt *et al* (2021)[1]	$1079.43	$313.34	$43.02	$202.35	$42.56	$130.51	$88.53	$70.04	$28.46	$443.65	$44.14	$24.68	N/A	$41.12	$44.35	$38.85

1. Costs were obtained between 2016-2019 for all studies within Alleweldt et al (2021), with the majority of costs from 2017, which was used for costing conversions to 2021 USD.

Table 12D. Output from Included Studies for the Scoping Review: Cost per Sample, Alleweldt et al. (2021)

Manuscript	Component Costs															
	WGS								**Conventional Method**							
	Staff (prof and tech): Avian influenza APHA	Staff (prof and tech): Avian influenza FLI	Staff (prof and tech): Influenza A+B EMC	Staff (prof and tech): Food-borne pathogens IZSLER	Staff (prof and tech): Food-borne pathogens ARG	Staff (prof and tech): Food-borne pathogens MDH	Staff (prof and tech): Food-borne pathogens PHAC	Staff (prof and tech): Food-borne pathogens PHE	Staff (prof and tech): Avian Influenza APHA	Staff (prof and tech): Avian Influenza FLI	Staff (prof and tech): influenza A+B EMC	Staff (prof and tech): Food-borne pathogens IZSLER	Staff (prof and tech): Food-borne pathogens ARG	Staff (prof and tech): Food-borne pathogens MDH	Staff (prof and tech): Food-borne pathogens PHAC	Staff (prof and tech): Food-borne pathogens PHE
Alleweldt *et al* (2021)[1]	$165.14	$126.37	$75.44	$81.10	€ 6.85	$25.73	$88.45	$46.05	$ 246.33	$415.22	$59.44	$35.98	N/A	$53.04	$59.69	$34.77

1. Costs were obtained between 2016-2019 for all studies within Alleweldt et al (2021), with the majority of costs from 2017, which was used for costing conversions to 2021 USD.

Table 12E. Output from Included Studies for the Scoping Review: Cost per Sample, Alleweldt et al. (2021)

Manuscript	Component Costs															
	WGS								**Conventional Method**							
	Other costs: Avian influenza APHA	Other costs: Avian influenza FLI	Other costs: Influenza A+B EMC	Other costs: Food-borne pathogens IZSLER	Other costs: Food-borne pathogens ARG	Other costs: Food-borne pathogens MDH	Other costs: Food-borne pathogens PHAC	Other costs: Food-borne pathogens PHE	Other costs: Avian Influenza APHA	Other costs: Avian Influenza FLI	Other costs: influenza A+B EMC	Other costs: Food-borne pathogens IZSLER	Other costs: Food-borne pathogens ARG	Other costs: Food-borne pathogens MDH	Other costs: Food-borne pathogens PHAC	Other costs: Food-borne pathogens PHE
Alleweldt *et al* (2021)[1]	$0.00	$0.00	$4.72	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$19.91	N/A	N/A	$0.00	$2.17

1. Costs were obtained between 2016-2019 for all studies within Alleweldt et al (2021), with the majority of costs from 2017, which was used for costing conversions to 2021 USD.

Figure 3. Quality of Evidence for Systematic Review: Modified CHEERS

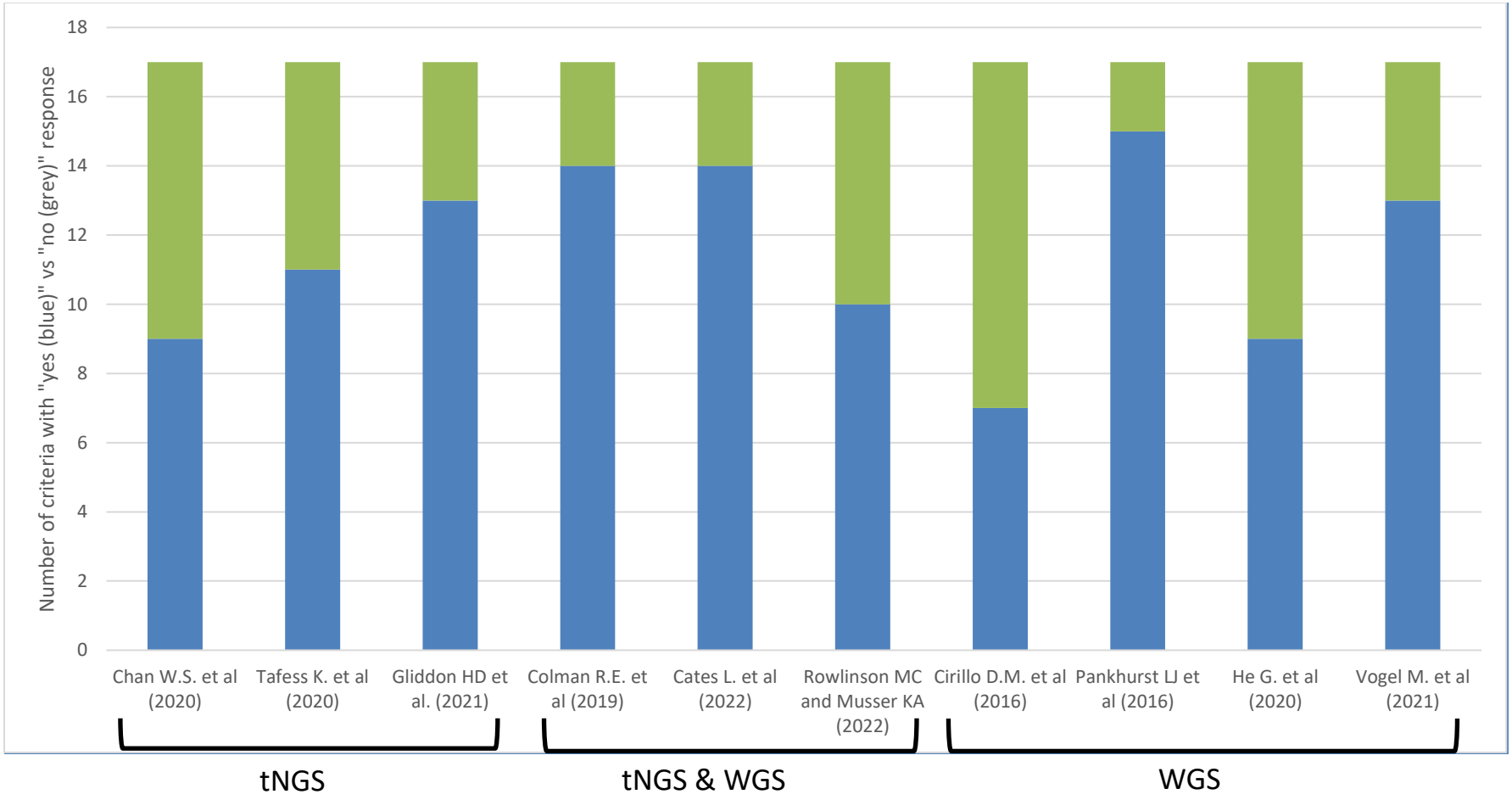
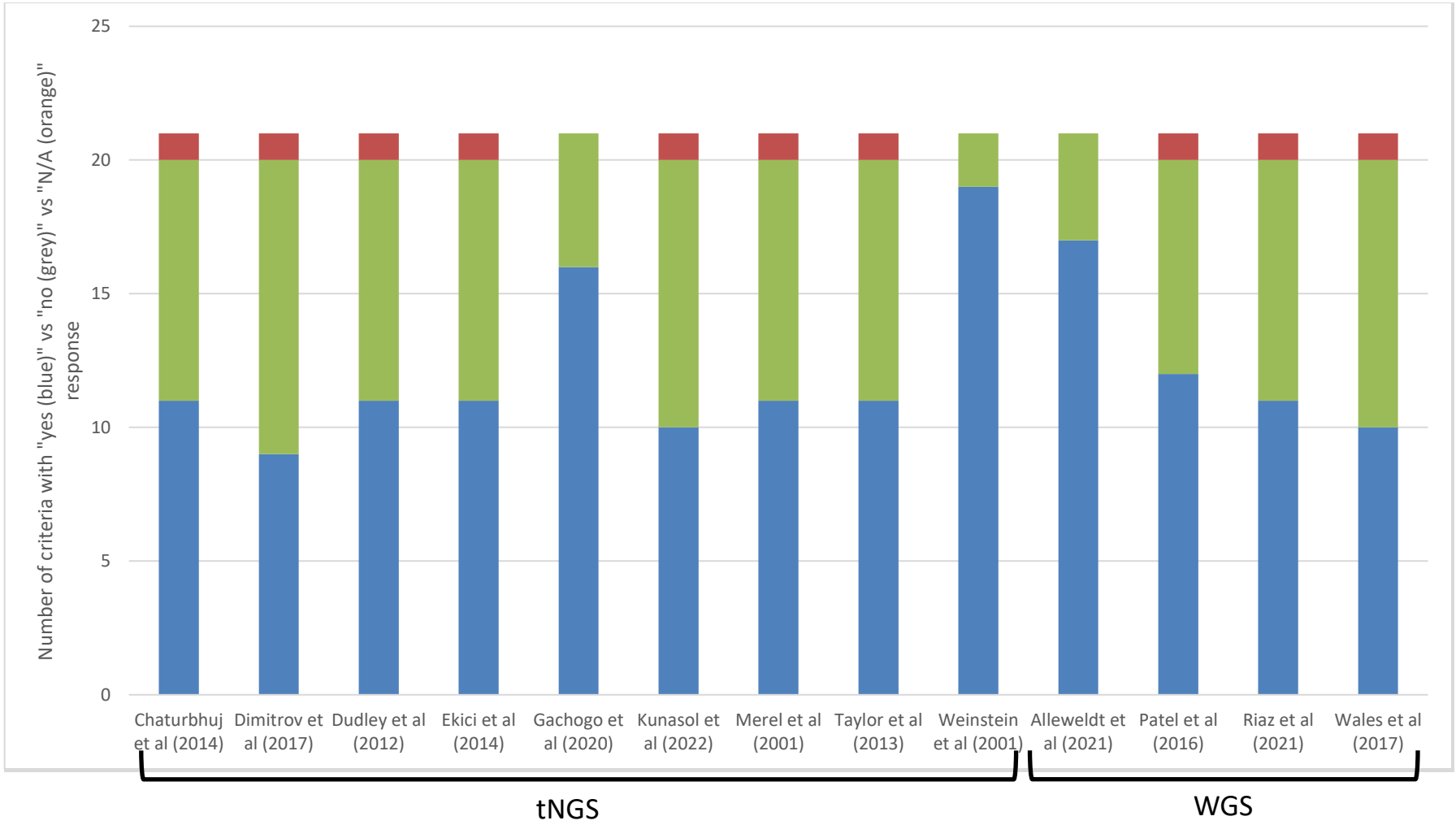


Figure 4. Quality of Evidence for Scoping Review: Modified CHEERS



7. REFERENCES

1) Koh WJ, Ko Y, Kim CK, Park KS, and Lee NY. Rapid diagnosis of tuberculosis and multidrug resistance using a MGIT 960 system. *Ann Lab Med.* 2012;32:264–9.
2) Zaheen A, and Bloom BR. Tuberculosis in 2020 - New Approaches to a Continuing Global Health Crisis. *N Engl J Med*. 2020 Apr 2;382(14):e26. doi: 10.1056/NEJMp2000325. PMID: 32242354.
3) Barbier M, Wirth T. The Evolutionary History, Demography, and Spread of the Mycobacterium tuberculosis Complex. Microbiol Spectr. 2016 Aug;4(4). doi: 10.1128/microbiolspec.TBTB2-0008-2016. PMID: 27726798.World Health Organization (WHO). *Global Tuberculosis Report 2022.* Available from: https://www.who.int/teams/global-tuberculosis-programme/tb-reports [Accessed 11th November 2022].
4) Viney K, Linh NW, and Gegia M, *et al*. New definitions of pre-extensively and extensively drug-resistant tuberculosis: update from the World Health Organization. *European Respiratory Journal.* 2021;57:1-5. Doi: 10.1183/13993003.00361-2021.
5) World Health Organization (WHO). *Tuberculosis fact sheet.* Available from: https://www.who.int/news-room/fact-sheets/detail/tuberculosis [Accessed 1st June 2022].
6) Dean AS, *et al* (2022). 25 years of surveillance of drug-resistant tuberculosis: achievements, challenges, and way forward. Lancet Infect Dis 22(7):E191-E196 (https://doi.org/10.1016/S1473-3099(21)00808-2).
7) Monedero-Recuero I, Gegia M, Wares DF, Chadha SS, and Mirzayev F. Situational analysis of 10 countries with a high burden of drug-resistant tuberculosis 2 years post-UNHLM declaration: progress and setbacks in a changing landscape. *Int J Infect Dis.* 2021 Jul;108:557-567. doi: 10.1016/j.ijid.2021.06.022. Epub 2021 Jun 15. PMID: 34139370.
8) Andama A, Somoskovi A, Mandel B, Bell D, and Gutierrez C. Improving diagnosis and case management of patients with tuberculosis: A review of gaps, needs and potential solutions in accessing laboratory diagnostics. *Infect Genet Evol*. 2019 Aug;72:131-140. doi: 10.1016/j.meegid.2018.08.030. Epub 2018 Oct 13. PMID: 30326281.
9) Furin J, Cox H, and Pai M. Tuberculosis. *Lancet*. 2019 Apr 20;393(10181):1642-1656. doi: 10.1016/S0140-6736(19)30308-3. Epub 2019 Mar 20. PMID: 30904262.
10) Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, and Hill-Cawthorne G, *et al*. Rapid determination of anti-tuberculosis drug resistance from whole genome sequences. Genome Med 2015;7:51.
11) Penn-Nicholson A, Gomathi SN, Ugarte-Gil C, Meaza A, Lavu E, Patel P, Choudhury B, Rodrigues C, Chadha S, Kazi M, Macé A, Nabeta P, Boehme C, Gangakhedkar RR, Sarin S, Tesfaye E, Gotuzzo E, du Cros P, Tripathy S, Ruhwald M, Singh M, Denkinger CM, Schumacher SG; Truenat Trial Consortium; and Members of the Truenat Trial Consortium. A prospective multicentre diagnostic accuracy study for the Truenat tuberculosis assays. *Eur Respir J*. 2021 Nov 4;58(5):2100526. doi: 10.1183/13993003.00526-2021. PMID: 34049948; PMCID: PMC8607906.
12) Jang JG, and Chung JH. (2020). Diagnosis and treatment of multidrug-resistant tuberculosis. *Yeungnam University journal of medicine*, *37*(4), 277–285. https://doi.org/10.12701/yujm.2020.00626.
13) World Health Organization. Molecular assays intended as initial tests for the diagnosis of pulmonary and extrapulmonary TB and rifampicin resistance in adults and children: rapid communication [Internet]. Geneva: World Health Organization; 2020 [cited 2020 Jun 5]. https://apps.who.int/iris/handle/10665/330395.
14) Ling DI, Zwerling A, and Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *European Respirology,* 2008;32:1165-1174.
15) Pai M, Nicol MP, and Boehme CC. Tuberculosis Diagnostics: State of the Art and Future Directions. *American Society for Microbiology.* 2016;4:5.

16) Cirillo DM, Miotto P, and Tortoli E. Evolution of Phenotypic and Molecular Drug Susceptibility Testing. *Adv Exp Med Biol.* 2017;1019:221-246. doi: 10.1007/978-3-319-64371-7_12. PMID: 29116638.
17) Hepple, *et al*. Implementation of liquid culture for tuberculosis diagnosis in a remote setting: lessons learned. *Int J Tuberc Lung Dis.* 2011;15(3):405–407.
18) World Health Organization. Molecular line probe assays for rapid screening of patients at risk of multidrug-resistant tuberculosis (MDR-TB) [Internet]. Geneva: World Health Organization; 2008 [cited 2020 Jul 14]. https://www.who.int/tb/laboratory/line_probe_assays/en/.
19) Lin M, Chen YW, Li YR, Long LJ, Qi LY, Cui TT, Wu SY, Lin JY, Wu T, Yang YC, Yuan WH, Wu GY, Lan QW, Liu JQ, Li YP, Yu ZY, and Guo XG. Systematic evaluation of line probe assays for the diagnosis of tuberculosis and drug-resistant tuberculosis. *Clin Chim Acta.* 2022 Aug 1;533:183-218. doi: 10.1016/j.cca.2022.06.020. Epub 2022 Jul 2. PMID: 35792161.
20) Theron G, Peter J, Richardson M, Barnard M, Donegan S, and Warren R, *et al*. The diagnostic accuracy of the GenoType(®) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs. *Cochrane Database Syst Rev* 2014; (10):CD010705.
21) Satta G, Lipman M, Smith GP, Arnold C, Kon OM, and McHugh TD. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? *Clin Microbiol Infect*. 2018 Jun;24(6):604-609. doi: 10.1016/j.cmi.2017.10.030. Epub 2017 Nov 3. PMID: 29108952.
22) Behjati S, and Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract* Ed. 2013 Dec;98(6):236-8. doi: 10.1136/archdischild-2013-304340. Epub 2013 Aug 28. PMID: 23986538; PMCID: PMC3841808.
23) Witney AA, Cosgrove CA, Arnold A, Hinds J, Stoker NG, and Butcher PD. Clinical use of whole genome sequencing for Mycobacterium tuberculosis. *BMC Med* 2016;14:46.
24) Bragg L, and Tyson GW. Metagenomics using next-generation sequencing. *Methods Mol Biol.* 2014;1096:183-201. doi: 10.1007/978-1-62703-712-9_15. PMID: 24515370.
25) Cabibbe AM, Spitaleri A, Battaglia S, Colman RE, Suresh A, Uplekar S, Rodwell TC, and Cirillo DM. Application of Targeted Next-Generation Sequencing Assay on a Portable Sequencing Platform for Culture-Free Detection of Drug-Resistant Tuberculosis from Clinical Samples. *J Clin Microbiol.* 2020 Sep 22;58(10):e00632-20. doi: 10.1128/JCM.00632-20. PMID: 32727827; PMCID: PMC7512157.
26) Moher D, Liberati A, Tetzlaff J, Altman DG, and PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ (Clinical research ed.)*, *339*, b2535. https://doi.org/10.1136/bmj.b2535.
27) Slatko BE, Gardner AF, and Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol.* 2018 Apr;122(1):e59. doi: 10.1002/cpmb.59. PMID: 29851291; PMCID: PMC6020069.
28) Colman RE, Anderson J, Lemmer D, Lehmkuhl E, Georghiou SB, Heaton H, Wiggins K, Gillece JD, Schupp JM, Catanzaro DG, Crudu V, Cohen T, Rodwell TC, aand Engelthaler DM. 2016. Rapid drug susceptibility testing of drug-resistant Mycobacterium tuberculosis isolates directly from clinical samples by use of amplicon sequencing: a proof-of-concept study. *J Clin Microbiol.* 54:2058–2067.
29) Mitchell SL, and Simner PJ. Next-Generation Sequencing in Clinical Microbiology: Are We There Yet? *Clin Lab Med*. 2019 Sep;39(3):405-418. doi: 10.1016/j.cll.2019.05.003. PMID: 31383265.
30) World Health Organization (WHO). *The use of next-generation sequencing technologies for the detection of mutations associated with rug resistance in Mycobacterium tuberculosis complex: technical guide.* Available from: https://apps.who.int/iris/bitstream/handle/10665/274443/WHO-CDS-TB-2018.19-eng.pdf?sequence=1&isAllowed=y [Accessed 20th June 2022].

31) Hemelaar J, Elangovan R, Yun J, Dickson-Tetteh L, Fleminger I, Kirtley S, Williams B, Gouws-Williams E, Ghys PD; WHO–UNAIDS Network for HIV Isolation Characterisation. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. Lancet Infect Dis. 2019 Feb;19(2):143-155. doi: 10.1016/S1473-3099(18)30647-9. Epub 2018 Nov 30. Erratum in: Lancet Infect Dis. 2020 Mar;20(3):e27. PMID: 30509777.
32) Bell LCK, Noursadeghi M. Pathogenesis of HIV-1 and Mycobacterium tuberculosis co-infection. Nat Rev Microbiol. 2018 Feb;16(2):80-90. doi: 10.1038/nrmicro.2017.128. Epub 2017 Nov 7. PMID: 29109555.
33) Michaud CM. Global Burden of Infectious Diseases. Encyclopedia of Microbiology. 2009:444–54. doi: 10.1016/B978-012373944-5.00185-1. Epub 2009 Feb 17. PMCID: PMC7150259.
34) Peteranderl C, Herold S, Schmoldt C. Human Influenza Virus Infections. Semin Respir Crit Care Med. 2016 Aug;37(4):487-500. doi: 10.1055/s-0036-1584801. Epub 2016 Aug 3. PMID: 27486731; PMCID: PMC7174870.
35) The National Institute for Health Research (NHS). *Guidance notes for registering a systematic review protocol with PROSPERO.* Available from: https://www.crd.york.ac.uk/prospero/documents/Registering%20a%20review%20on%20PROSPERO.pdf [Accessed 22nd June 2022].
36) International Monetary Fund (IMF). *Inflation rate, average consumer prices.* Available from: https://www.imf.org/external/datamapper/PCPIPCH@WEO/OEMDC [Accessed 12th December 2022].
37) Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E, and ISPOR Health Economic Evaluation Publication Guidelines-CHEERS Good Reporting Practices Task Force (2013). Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *16*(2), 231–250. https://doi.org/10.1016/j.jval.2013.02.002.
38) Tafess K, Ng TTL, Lao HY, Leung KSS, Tam KKG, Rajwani R, Tam STY, Ho LPK, Chu CMK, Gonzalez D, Sayada C, Ma OCK, Nega BH, Ameni G, Yam WC, and Siu GKH. Targeted-Sequencing Workflows for Comprehensive Drug Resistance Profiling of Mycobacterium tuberculosis Cultures Using Two Commercial Sequencing Platforms: Comparison of Analytical and Diagnostic Performance, Turnaround Time, and Cost. *Clin Chem*. 2020 Jun 1;66(6):809-820. doi: 10.1093/clinchem/hvaa092. PMID: 32402055.
39) Cates L, Codreanu A, Ciobanu N, Fosburgh H, Allender CJ, Centner H, Engelthaler DM, Crudu V, Cohen T, and Menzies NA. Budget impact of next-generation sequencing for diagnosis of TB drug resistance in Moldova. *Int J Tuberc Lung Dis*. 2022 Oct 1;26(10):963-969. doi: 10.5588/ijtld.22.0104. PMID: 36163669.
40) Colman RE, Mace A, Seifert M, Hetzel J, Mshaiel H, Suresh A, Lemmer D, Engelthaler DM, Catanzaro DG, Young AG, Denkinger CM, and Rodwell TC. Whole-genome and targeted sequencing of drug-resistant Mycobacterium tuberculosis on the iSeq100 and MiSeq: A performance, ease-of-use, and cost evaluation. *PLoS Med.* 2019 Apr 30;16(4):e1002794. doi: 10.1371/journal.pmed.1002794.
41) Rowlinson MC, and Musser KA. Current Methods and Role of Next-Generation Sequencing in the Diagnosis of Antimicrobial Resistance in Tuberculosis. *Clinical Microbiology Newsletter.* 2022 Jan;44:1-12.doi: https://doi.org/10.1016/j.clinmicnews.2021.12.001.
42) Chan WS, Au CH, Chung Y, Leung HCM, Ho DN, Wong EYL, Lam TW, Chan TL, Ma ESK, and Tang BSF. Rapid and economical drug resistance profiling with Nanopore MinION for clinical specimens with low bacillary burden of Mycobacterium tuberculosis.

BMC Res Notes. 2020 Sep 18;13(1):444. doi: 10.1186/s13104-020-05287-9. PMID: 32948225; PMCID: PMC7501614.
43) Gliddon HD, Frampton D, Munsamy V, Heaney J, Pataillot-Meakin T, Nastouli E, Pym AS, Steyn AJC, Pillay D, and McKendry RA. A Rapid Drug Resistance Genotyping Workflow for Mycobacterium tuberculosis, Using Targeted Isothermal Amplification and Nanopore Sequencing. *Microbiol Spectr.* 2021 Dec 22;9(3):e0061021. doi: 10.1128/Spectrum.00610-21. Epub 2021 Nov 24. PMID: 34817282; PMCID: PMC8612157.
44) He G, Li Y, Chen X, Chen J, and Zhang W. Prediction of treatment outcomes for multidrug-resistant tuberculosis by whole-genome sequencing. *Int J Infect Dis*. 2020 Jul;96:68-72. doi: 10.1016/j.ijid.2020.04.043. Epub 2020 Apr 24. PMID: 32339719.
45) Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW; and COMPASS-TB Study Group. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med.* 2016 Jan;4(1):49-58. doi: 10.1016/S2213-2600(15)00466-X. Epub 2015 Dec 4. PMID: 26669893; PMCID: PMC4698465.
46) Vogel M, Utpatel C, Corbett C, Kohl TA, Iskakova A, Ahmedov S, Antonenka U, Dreyer V, Ibrahimova A, Kamarli C, Kosimova D, Mohr V, Sahalchyk E, Sydykova M, Umetalieva N, Kadyrov A, Kalmambetova G, Niemann S, and Hoffmann H. Implementation of whole genome sequencing for tuberculosis diagnostics in a low-middle income, high MDR-TB burden country. *Sci Rep.* 2021 Jul 28;11(1):15333. doi: 10.1038/s41598-021-94297-z. PMID: 34321545; PMCID: PMC8319420.
47) Alleweldt F, Kara Ş, Best K, Aarestrup FM, Beer M, Bestebroer TM, Campos J, Casadei G, Chinen I, Van Domselaar G, Dominguez C, Everett HE, Fouchier RA, Grant K, Green J, Höper D, Johnston J, Koopmans MP, Oude Munnink BB, Myers R, Nadon C, Patel A, Pohlmann A, Pongolini S, Reimer A, Thiessen S, Wylezich C. Economic evaluation of whole genome sequencing for pathogen identification and surveillance - results of case studies in Europe and the Americas 2016 to 2019. Euro Surveill. 2021 Mar;26(9):1900606. doi: 10.2807/1560-7917.ES.2021.26.9.1900606. PMID: 33663647; PMCID: PMC7934224.
48) Dimitrov KM, Sharma P, Volkening JD, Goraichuk IV, Wajid A, Rehmani SF, Basharat A, Shittu I, Joannis TM, Miller PJ, Afonso CL. A robust and cost-effective approach to sequence and analyze complete genomes of small RNA viruses. Virol J. 2017 Apr 7;14(1):72. doi: 10.1186/s12985-017-0741-5. PMID: 28388925; PMCID: PMC5384157.
49) Weinstein MC, Goldie SJ, Losina E, Cohen CJ, Baxter JD, Zhang H, Kimmel AD, Freedberg KA. Use of genotypic resistance testing to guide hiv therapy: clinical impact and cost-effectiveness. Ann Intern Med. 2001 Mar 20;134(6):440-50. doi: 10.7326/0003-4819-134-6-200103200-00008. PMID: 11255519.
50) Gachogo RW, Mwai DN, Onyambu FG. Cost analysis of implementing HIV drug resistance testing in Kenya: a case study of a service delivery site at a tertiary level hospital in Kenya. F1000Res. 2020 Jul 29;9:793. doi: 10.12688/f1000research.23379.1. PMID: 32983418; PMCID: PMC7495211.
51) Riaz N, Leung P, Barton K, Smith MA, Carswell S, Bull R, Lloyd AR, Rodrigo C. Adaptation of Oxford Nanopore technology for hepatitis C whole genome sequencing and identification of within-host viral variants. BMC Genomics. 2021 Mar 2;22(1):148. doi: 10.1186/s12864-021-07460-1. PMID: 33653280; PMCID: PMC7923462.
52) Merel P, Pellegrin I, Garrigue I, Caumont A, Schrive MH, Birac V, Bonot P, Fleury H. Comparison of capillary electrophoresis sequencing with the new CEQ 2000 DNA Analysis System to conventional gel based systems for HIV drug resistance analysis. J Virol Methods. 2001 Oct;98(1):9-16. doi: 10.1016/s0166-0934(01)00338-x. PMID: 11543879.
53) Chaturbhuj DN, Nirmalkar AP, Paranjape RS, Tripathy SP. Evaluation of a cost effective in-house method for HIV-1 drug resistance genotyping using plasma samples. PLoS One.

2014 Feb 12;9(2):e87441. doi: 10.1371/journal.pone.0087441. PMID: 24533056; PMCID: PMC3922725.
54) Dudley DM, Chin EN, Bimber BN, Sanabani SS, Tarosso LF, Costa PR, Sauer MM, Kallas EG, O'Connor DH. Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. PLoS One. 2012;7(5):e36494. doi: 10.1371/journal.pone.0036494. Epub 2012 May 4. PMID: 22574170; PMCID: PMC3344889.
55) Ekici H, Rao SD, Sönnerborg A, Ramprasad VL, Gupta R, Neogi U. Cost-efficient HIV-1 drug resistance surveillance using multiplexed high-throughput amplicon sequencing: implications for use in low- and middle-income countries. J Antimicrob Chemother. 2014 Dec;69(12):3349-55. doi: 10.1093/jac/dku278. Epub 2014 Jul 31. PMID: 25085657.
56) Kunasol C, Dondorp AM, Batty EM, Nakhonsri V, Sinjanakhom P, Day NPJ, Imwong M. Comparative analysis of targeted next-generation sequencing for Plasmodium falciparum drug resistance markers. Sci Rep. 2022 Apr 1;12(1):5563. doi: 10.1038/s41598-022-09474-5. PMID: 35365711; PMCID: PMC8974807.
57) Patel N, Ferns BR, Nastouli E, Kozlakidis Z, Kellam P, and Morris S. Cost analysis of standard Sanger sequencing versus next generation sequencing in the ICONIC study. The Lancet Meeting Abstracts. 2016 Nov 25.
58) Taylor SM, Parobek CM, Aragam N, Ngasala BE, Mårtensson A, Meshnick SR, Juliano JJ. Pooled deep sequencing of Plasmodium falciparum isolates: an efficient and scalable tool to quantify prevailing malaria drug-resistance genotypes. J Infect Dis. 2013 Dec 15;208(12):1998-2006. doi: 10.1093/infdis/jit392. Epub 2013 Aug 1. PMID: 23908494; PMCID: PMC3836461.
59) Wales C, Corpus G, Chang C. Hepatitis C Virus Genotyping Using Next-Generation Sequencing: An Efficient Alternative to Sanger Sequencing. Arch Pathol Lab Med. 2017 Mar;141(3):323-324. doi: 10.5858/arpa.2016-0338-LE. PMID: 28234569.
60) Mugwagwa T, Abubakar I, White PJ. Using molecular testing and whole-genome sequencing for tuberculosis diagnosis in a low-burden setting: a cost-effectiveness analysis using transmission-dynamic modelling. Thorax. 2019;0:1-11. doi:10.1136/thoraxjnl-2019-214004.

8. APPENDIX

8.1 Results of Economic Evidence around using WGS for the Diagnosis of DR-TB

Seven manuscripts looked at cost data for WGS (Tables 2A-B). Two manuscripts were from the US, one in countries across Europe (UK, Ireland, Germany and France) as well as Canada, one in Italy, one in the Kyrgyz Republic, and one in China (Tables 2A-B). The year of cost valuation ranged from 2014-2021 across studies (Tables 2A-B). Six manuscripts assessed diagnostic cost per sample for WGS, and two assessed this for conventional pDST (Tables 2A-B). Four of the manuscripts considered component costs for WGS, and two for pDST as a diagnostic comparator. The manuscript by Vogel M *et al.* (2021) was the only to consider cost per step of WGS (39). Similar with the tNGS papers, the study by Cates *et al.* (2022) was the only one to perform a BIA for WGS (39). Of the WGS papers, only the study by Colman *et al. (*2019) was in-house, and the rest were performed centrally (Tables 2A-B).

For WGS, the diagnostic cost per sample was $68.00 using the Illumina NextSeq platform (n=1), between $71.61 and $277.00 for Illumina MiSeq (n=6), and $63.00 for Nanopore MinION (n=1) (Tables 10A). The manuscript by Cirillo *et al.* (2016) reported a diagnostic cost per sample of $185.77 in 2016 using the Illumina MiSeq platform, which included labor costs, reagents and consumables, as well as equipment (41). The large range for diagnostic costs per sample using the MiSeq platform was a result of a sensitivity analysis where authors investigated different sample volumes finding that increasing sample volumes by 10% decreased per sample cost ($109.67) and decreasing samples by 10% increased per sample cost ($112.38); using 600- ($141.00) as opposed to 500-cycles ($167.00) reduced costs through optimizing the same flow cell, which is a major cost item; and the initial samples ($277.00 per sample for initial 174 samples) were more expensive than the subsequent samples ($141.00-$167.00 per sample for the remaining samples), due to one-time costs associate with training and errors as the sequencing protocol was still being established (Tables 10A-C). The manuscript by Vogel *et al.* (2021) was the only paper to assess the costs per step of WGS for each sample, and looked at this using the Illumina MiSeq platform (39). Overall, the step with the lowest costs was target enrichment multiplex PCR ($1.31) and the step with the greatest costs was sequencing (up to $156.32) (39).

8.2. Systematic Review Search Strategy

PubMed search

Concept #1

"Tuberculosis, Multidrug-Resistant"[Mesh] OR "Extensively Drug-Resistant Tuberculosis"[Mesh] OR "Multidrug-Resistant Tuberculos*"[tw] OR "MDR Tuberculos*"[tw] OR "Extensively Drug Resistant Tuberculos*"[tw] OR "Extremely Drug Resistant Tuberculos*"[tw] OR "XDR-TB"[tw] "XDRTB"[tw] OR "multi-drug resistant TB"[tw] OR "multidrug resistant TB"[tw] OR "multiresistant tuberculos*"[tw] OR "extensively drug resistant TB"[tw] OR "extensively drug-resistant tuberculos*"[tw] OR "XDR-TB"[tw] OR "XDRTB"[tw] OR "drug resistant TB"[tw] OR "resistant pulmonary

TB"[tw] OR "resistant pulmonary tuberculos*"[tw] OR "resistant TB"[tw] OR "resistant tuberculos*"[tw]

Concept #2

"Costs and Cost Analysis"[mesh] OR "Costs and Cost Analys*"[tw] OR "Cost Analys*"[tw] OR "Cost Benefit Analys*"[tw] OR "Cost Effectiveness"[tw] OR "economic*"[tw] OR "cost allocation"[tw] OR "cost efficiency analys*"[tw]

Concept #3

"Sequence Analysis"[mesh] OR "Sequence Analys*"[tw] OR "Sequence Determination*"[tw] OR "next-generation sequencing"[tw] OR " next-gen sequence analys*"[tw] OR "next-gen sequencing"[tw] OR "genetic sequencing"[tw] OR "genomic sequencing"[tw]

#1 AND #2 AND #3 – 17 records

Embase search

Concept #1

'multidrug resistant tuberculosis'/exp OR 'drug resistant tuberculosis'/exp OR 'extensively drug resistant tuberculosis'/exp OR ('MDR-TB' OR 'multidrug-resistant tuberculos*' OR 'multi-drug resistant tuberculos*' OR 'multi-resistant tuberculos*' OR 'multi-drug resistant TB' OR 'multidrug resistant TB' OR 'multiresistant tuberculos*' OR 'extensively drug resistant TB' OR 'extensively drug-resistant tuberculos*' OR 'XDR-TB' OR 'XDRTB' OR 'drug resistant TB' OR 'resistant pulmonary TB' OR 'resistant pulmonary tuberculos*' OR 'resistant TB' OR 'resistant tuberculos*'):ab,ti,kw

Concept #2

'cost'/exp OR 'cost effectiveness analysis'/exp OR 'cost benefit analysis'/exp OR ('cost analys*' OR 'cost benefit' OR 'cost-benefit analys*' OR 'cost effectiveness' OR 'cost efficiency analys*' OR 'cost allocation'):ab,ti,kw

Concept #3

'sequence analysis'/exp OR 'high throughput sequencing'/exp OR ('high through-put sequencing' OR 'Sequence Analys*' OR 'Sequence Determination*' OR 'next-generation sequencing' OR ' genetic sequencing' OR 'genomic sequencing'):ab,ti,kw

#1 AND #2 AND #3 – 26 records

Scopus search

Concept #1

"multidrug resistant tuberculosis" OR "drug resistant tuberculosis" OR "extensively drug resistant tuberculosis" OR "MDR-TB" OR "multidrug-resistant tuberculos*" OR "multi-drug resistant tuberculos*" OR "multi-resistant tuberculos*" OR "multi-drug resistant TB" OR "multidrug resistant TB" OR "multiresistant tuberculos*" OR "extensively drug resistant TB" OR "extensively drug-resistant tuberculos*" OR "XDR-TB" OR "XDRTB" OR "drug resistant TB" OR "resistant pulmonary TB" OR "resistant pulmonary tuberculos*" OR "resistant TB" OR "resistant tuberculos*"

Concept #2

"cost" OR "cost effectiveness analys*" OR "cost analys*" OR "cost benefit" OR "cost-benefit analys*" OR "cost effectiveness" OR "cost efficiency analys*" OR "cost allocation"

Concept #3

"sequence analys* " OR "high throughput sequencing" OR "Sequence Determination*" OR "next-generation sequencing" OR "genetic sequencing" OR "genomic sequencing"
#1 AND #2 AND #3 – 56 records

8.3. Scoping Review Search Strategy

PubMed search

Concept #1A

"Hepatitis"[Mesh:NoExp] OR "Hepatitis, Viral, Human"[Mesh] OR "hepatitis"[tiab]

Concept #1B

"HIV"[mesh] OR "HIV"[tw] OR "Human Immunodeficiency Virus*"[tw] OR "AIDS Virus*"[tw] OR "Acquired Immune Deficiency Syndrome Virus"[tw] OR "Acquired Immunodeficiency Syndrome Virus"[tw]

Concept #1C

("Malaria"[mesh] OR "Malaria"[tw]) OR ("Plasmodium Infection*"[tw] OR "Remittent Fever"[tw] OR "Marsh Fever"[tw] "malarial fever"[tw] OR "malarial infection*"[tw] OR "paludism"[tw] OR "Plasmodia infection*"[tw] OR "plasmodial infection*"[tw] OR "plasmodiosis"[tw])

Concept #1D

"Influenza, human"[mesh] OR "Influenza*"[tw] OR "Human Flu"[tw] OR "Grippe"[tw]

Concept #2

"Costs and Cost Analysis"[mesh] OR "Costs and Cost Analys*"[tw] OR "Cost Analys*"[tw] OR "Cost Benefit Analys*"[tw] OR "Cost Effectiveness"[tw] OR "economic*"[tw] OR "cost allocation"[tw] OR "cost efficiency analys*"[tw]

Concept #3

"Sequence Analysis"[mesh] OR "Sequence Analys*"[tw] OR "Sequence Determination*"[tw] OR "next-generation sequencing"[tw] OR " next-gen sequence analys*"[tw] OR "next-gen sequencing"[tw] OR "genetic sequencing"[tw] OR "genomic sequencing"[tw]

#1A AND #2 AND #3 – 63 records
#1B AND #2 AND #3 – 62 records
#1C AND #2 AND #3 – 29 records
#1D AND #2 AND #3 – 79 records

Embase search

Concept #1A

'virus hepatitis'/exp OR 'hepatitis'/mj OR ('viral hepatitis' OR 'viral hepatitis'):ab,ti,kw

Concept #1B

'Human immunodeficiency virus'/exp OR ('HIV' OR 'Human Immunodeficiency Virus*' OR 'AIDS Virus*' OR 'Acquired Immune Deficiency Syndrome Virus' OR 'Acquired Immunodeficiency Syndrome Virus'):ab,ti,kw

Concept #1C

'Malaria'/exp OR ('Malaria' OR 'Plasmodium Infection*' OR 'Remittent Fever' OR 'Marsh Fever' OR 'malarial fever' OR 'malarial infection*' OR 'marsh fever' OR 'paludism' OR 'Plasmodia infection*' OR 'plasmodial infection*' OR 'plasmodiosis'):ab,ti,kw

Concept #1D

'influenza'/exp OR ('Influenza*' OR 'Human Flu' OR 'Grippe'):ab,ti,kw

Concept #2

'cost'/exp OR 'cost effectiveness analysis'/exp OR 'cost benefit analysis'/exp OR ('cost analys*' OR 'cost benefit' OR 'cost-benefit analys*' OR 'cost effectiveness' OR 'cost efficiency analys*' OR 'cost allocation'):ab,ti,kw

Concept #3

'sequence analysis'/exp OR 'high throughput sequencing'/exp OR ('high through-put sequencing' OR 'Sequence Analys*' OR 'Sequence Determination*' OR 'next-generation sequencing' OR ' genetic sequencing' OR 'genomic sequencing'):ab,ti,kw

#1A AND #2 AND #3 – 42 records
#1B AND #2 AND #3 – 76 records
#1C AND #2 AND #3 – 19 records
#1D AND #2 AND #3 – 38 records

Scopus search

Concept #1A

"virus hepatitis " OR "viral hepatitis" OR "hepatitis"

Concept #1B

"Human immunodeficiency virus* " OR "HIV" OR "AIDS Virus* " OR "Acquired Immune Deficiency Syndrome Virus" OR "Acquired Immunodeficiency Syndrome Virus"

Concept #1C

"Malaria" OR "Plasmodium Infection*" OR "Remittent Fever" OR "Marsh Fever" OR "malarial fever" OR "malarial infection*" OR "marsh fever" OR "paludism " OR "Plasmodia infection*" OR "plasmodial infection*" OR "plasmodiosis"

Concept #1D

Influenza OR "Human Flu" OR "Grippe"

Concept #2

"cost" OR "cost effectiveness analys*" OR "cost analys*" OR "cost benefit" OR "cost-benefit analys*" OR "cost effectiveness" OR "cost efficiency analys*" OR "cost allocation"

Concept #3

"sequence analys* " OR "high throughput sequencing" OR "Sequence Determination*" OR "next-generation sequencing" OR "genetic sequencing" OR "genomic sequencing"

#1A AND #2 AND #3 – 161 records
#1B AND #2 AND #3 – 277 records
#1C AND #2 AND #3 – 98 records
#1D AND #2 AND #3 – 98 records

Supplementary Table 1A. Quality of Health Economic Studies Instrument Assessment of Included tNGS Studies for the Systematic Review

Manuscript	Was the study population clearly described?	Were competing alternatives clearly described?	Was a well-defined research question posed in answerable form?	Was the actual perspective chosen appropriate?	Were all important and relevant costs for each alternative identified?	Were all costs measured appropriately in physical units?	Were costs valued appropriately?	Were all outcomes measured appropriately in physical units?
Colman RE et al (2019)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Chan WS et al (2020)	No	Yes	Yes	Yes	No	Yes	No	Yes
Tafess K et al (2020)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Gliddon HD et al (2021)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Cates L et al (2022)	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Rowlinson MC and Musser KA (2022)[1]	No	Yes	Yes	Yes	No	Yes	Yes	Yes

1. The paper by Rowlinson MC and Musser KA (2020) is a report.

Supplementary Table 1B. Quality of Health Economic Studies Instrument Assessment of Included tNGS Studies for the Systematic Review

Manuscript	Was any sensitivity analysis performed around uncertain values?	Did the conclusions follow from the data reported?	Were component costs reported?	Were the sources of all costing included?	Did the authors describe methods for converting costs into a common currency base and the exchange rate?	Did the authors state the time horizon(s) over which costs were being evaluated and say why appropriate?	Did the study discuss the generalizability of the results to other settings and patient/client groups?	Did the article indicate that there was no potential conflict of interest of study researcher(s) and funder(s)?	Were ethical and distributional issues discussed appropriately?
Colman RE et al (2019)	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Chan WS et al (2020)	No	No	No	Yes	No	No	Yes	Yes	Yes
Tafess K et al (2020)	No	Yes	No	No	Yes	No	No	Yes	Yes
Gliddon HD et al (2021)	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Cates L et al (2022)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes

Rowlinson MC and Musser KA (2022)[1]	No	Yes	No	Yes	No	No	Yes	No	Yes

1. The paper by Rowlinson MC and Musser KA (2020) is a report.

Supplementary Table 2A. Quality of Health Economic Studies Instrument Assessment of Included WGS Studies for the Systematic Review

Manuscript	Was the study population clearly described?	Were competing alternatives clearly described?	Was a well-defined research question posed in answerable form?	Was the actual perspective chosen appropriate?	Were all important and relevant costs for each alternative identified?	Were all costs measured appropriately in physical units?	Were costs valued appropriately?	Were all outcomes measured appropriately in physical units?
Cirillo DM et al (2016)	No	No	Yes	Yes	No	Yes	No	Yes
Pankhurst LJ et al (2016)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Colman RE et al (2019)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
He G et al (2020)	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Vogel M et al (2021)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rowlinson MC and Musser KA (2022)[1]	No	Yes	Yes	Yes	No	Yes	Yes	Yes

1. The paper by Rowlinson MC and Musser KA (2020) is a report.

Supplementary Table 2B. Quality of Health Economic Studies Instrument Assessment of Included WGS Studies for the Systematic Review

Manuscript	Was any sensitivity analysis performed around uncertain values?	Did the conclusions follow from the data reported?	Were component costs reported?	Were the sources of all costing included?	Did the authors describe methods for converting costs into a common currency base and the exchange rate?	Did the authors state the time horizon(s) over which costs were being evaluated and say why appropriate?	Did the study discuss the generalizability of the results to other settings and patient/client groups?	Did the article indicate that there was no potential conflict of interest of study researcher(s) and funder(s)?	Were ethical and distributional issues discussed appropriately?
Cirillo DM et al (2016)	No	Yes	No	No	No	Yes	No	Yes	No
Pankhurst LJ et al (2016)	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Colman RE et al (2019)	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
He G et al (2020)	No	Yes	No	No	No	No	No	Yes	Yes
Vogel M et al (2021)	No	Yes	Yes	No	No	No	Yes	Yes	Yes

Rowlinson MC and Musser KA (2022)[1]	No	Yes	No	Yes	No	No	Yes	No	Yes

1. The paper by Rowlinson MC and Musser KA (2020) is a report

Supplementary Table 3A. Quality of Health Economic Studies Instrument Assessment of Included Studies for the Scoping Review

Manuscript	Was the study population clearly described?	Were competing alternatives clearly described?	Was a well-defined research question posed in answerable form?	Was the actual perspective chosen appropriate?	Were all important and relevant costs for each alternative identified?	Were all costs measured appropriately in physical units?	Were costs valued appropriately?	Were all outcomes measured appropriately in physical units?	Was any sensitivity analysis performed around uncertain values?	Did the conclusions follow from the data reported?	Were component costs reported?
Alleweldt et al (2021)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Chaturbhuj et al (2014)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Dimitrov et al (2017)	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Dudley et al (2012)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Ekici et al (2014)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Gachogo et al (2020)	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Kunasol et al (2022)	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No
Merel et al (2001)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Patel et al (2016)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Riaz et al (2021)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Taylor et al (2013)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Wales et al (2017)	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
Weinstein et al (2001)	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No

Supplementary Table 3B. Quality of Health Economic Studies Instrument Assessment of Included Studies for the Scoping Review

Manuscript	Were the sources of all costing included?	Did the authors describe methods for converting costs into a common currency base and the exchange rate?	Did the authors state the time horizon(s) over which costs were being evaluated and say why appropriate?	Did the study discuss the generalizability of the results to other settings and patient/client groups?	Did the article indicate that there was no potential conflict of interest of study researcher(s) and funder(s)?	Was the economic study design appropriate to the stated objective?	Was an incremental analysis of costs and outcomes of alternatives performed?	Were all future costs and outcomes discounted appropriately?	Was the comparator stated clearly and was it appropriate?	Was the chosen time horizon appropriate in order to include relevant costs and consequences?
Alleweldt et al (2021)	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Chaturbhuj et al (2014)	No	No	No	No	Yes	Yes	No	No	Yes	N/A
Dimitrov et al (2017)	No	No	No	No	Yes	Yes	No	No	No	N/A
Dudley et al (2012)	No	No	No	No	Yes	Yes	No	No	Yes	N/A
Ekici et al (2014)	No	No	No	Yes	No	Yes	No	No	Yes	N/A
Gachogo et al (2020)	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes
Kunasol et al (2022)	No	No	No	No	Yes	Yes	No	No	Yes	N/A
Merel et al (2001)	Yes	No	No	No	No	Yes	No	No	Yes	N/A
Patel et al (2016)	No	No	Yes	No	Yes	Yes	No	No	Yes	N/A
Riaz et al (2021)	No	No	No	No	Yes	Yes	No	No	Yes	N/A
Taylor et al (2013)	No	No	No	No	Yes	Yes	No	No	Yes	N/A
Wales et al (2017)	No	No	No	No	Yes	Yes	No	No	Yes	N/A
Weinstein et al (2001)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

9. ACKNOWLEDGEMENTS

There are a number of individuals whom we wish to thank for their support in completing this work. Firstly, we would like to thank Donna Hesson for her help with designing our search strategy and for being an altogether great resource. Donna is one of the Public Health Informationists at Johns Hopkins University, Bloomberg School of Public Health. Secondly, we want to thank several members of FIND, for providing their technical expertise, as well as for making themselves available to consult whenever we had questions. Thirdly, we want to thank the authors from several of the manuscripts that were included in our systematic review. Our conversations were instrumental in better understanding their methodology, and some of the cost considerations with tNGS.

Web Annex D.20. Stakeholder experiences and perspectives of using Targeted next generation sequencing for diagnosing MDR-TB: interviews results

(Draft 1, 15 December 2022. Final draft 17 April 2023)

Author:
Dr Natalie Leon
Independent public health research consultant
nataliehelenleon@gmail.com

A project for:
Unit: TB Prevention, Diagnosis, Treatment, Care & Innovation (PCI)

Abstract

BACKGROUND

Timely drug resistance detection is essential to global tuberculosis management. Novel molecular testing tools are being developed to rapidly detect multi-drug resistant Tuberculosis (MDR-TB) which is needed for more effective and timely treatment of MDR-TB. Targeted next generation sequencing (tNGS), is new molecular-based rapid diagnostic technology (used on a sputum sample), that can provide comprehensive diagnosis of MDR-TB in a short turnaround time of a few days, compared to the standard culture-based phenotypic MDR-TB diagnostics that take multiple weeks. There are, however, obstacles to wide-scale and effective implementation of rapid diagnostics, that relate to feasibility and other contextual factors, and there is a need to better understand these barriers. The experiences and perceptions of health workers implementing rapid TB diagnostics can provide useful information to better understand the implementation factors that may shape the use and impact of rapid diagnostics.

OBJECTIVE

This study provides evidence to inform the formation of a new WHO guideline for use of tNGS technology for MDR-TB diagnosis. The objective is to explore the perceptions and experiences of implementers of tNGS technology, with respect to acceptability, feasibility, values and preferences and equity.

METHODS

This was a cross-sectional observational qualitative study using semi-structured interviews. Laboratory staff and management who were involved with testing tNGS platforms in the three FIND trial sites, India, Georgia, and South Africa, were the main participants. Other stakeholders were included to gain a broader perspective. Fifteen (15) interviews were conducted with 17 individuals, during September to October 2022. Data was analysed by identifying and synthesizing the main themes that emerged.

RESULTS

Acceptability

There was an overwhelmingly positive sentiment for the acceptability and potential utility of tNGS technology. tNGS was seen as a 'major advancement' in molecularly MDR TB diagnostics.

1. The main reasons for the **high level of acceptability were the comprehensiveness** (resistance diagnosis for more drugs and for newest and repurposed drugs), **the convenience of using sputum sample** (as compared to culture samples), **and the rapidness** (quick results compared to phenotypic testing times; 3-5 days as compared to 4-6 weeks).
2. There was also the sense that there is **a good window of opportunity for the utility of tNGS technology;** that the technology is arriving at the right time given that resistance to newer TB drugs is likely to increase as use if these drugs become routine.

Feasibility

Though there was high praise for the capability and potential utility of tNGS technology, **there were several challenges identified when testing the tNGS platforms.** The sense was that these challenges limited the current feasibility of tNGS for routine uptake. The overall sentiment was that the tNGS technology needs to be further developed before it can be considered fully ready for operational use. The following feasibility challenges were identified:

1. **Start-up and setting up challenges:** There were multiple starting and setting up problems. Some related to the newness of the technology and the trial setting, problems with importing technology and specialist supplies, problems related to absence of in-country technical assistance for problem-solving, as well as need for more hands-on training practice.
2. **High technical complexity of the test is a challenge:** tNGS technology was viewed as a high complexity molecular test that was technically challenging. For example, preparing the sample for sequencing involves multiple steps, that require attention to detail, precision, and with little room for error. The complexity of the library preparation phase was more particular for the Deeplex platform, though both the Deeplex and the Nanopore platforms were thought to have different pros and cons in terms of complexity. Both platforms were thought to have insufficient opportunities for early error recognition and error correction, and this increased the risk of failed runs.

3. **Specialized laboratory infrastructure and human resources are required which are potentially challenging**: As tNGS is a molecular-based testing platform, the platform requires highly specialised laboratory infrastructure that includes multiple rooms to prevent contamination and specialized cold storage facilities. Highly specialized molecular/medical scientists are needed to perform the tests. In these LMIC settings, such specialized laboratory infrastructure and staff may only be available at centralized laboratories and not necessarily at regional laboratories.
4. **Specialist requirements for operating the test are potentially challenging:** In addition to highly specialized laboratory infrastructure and staff, the testing technology also requires uninterrupted supply of electricity, high internet connectivity, high computer capacity, clean water, and temperature controls - requirements that may pose challenges in some LMIC settings.
5. **Supply chain challenges was an obstacle:** A major concern was the supply chain challenges that were encountered. Procurement bottle-necks and delays jeopardized continuous access to specialist supplies.
6. **Data management and storage requirements presented challenges:** There were concerns that data analysis and data storage requirements were not fully developed, including systems for backing up data, data ownership and data security considerations. Consideration is needed for how tNGS and routine laboratory information systems would be interlinked.
7. **Continuous updating of the WHO mutations reference library would be required:** There is the sense that the usefulness of the tNGS technology is dependent on the informational support provided by the WHO mutations reference library, which allows for meaningful interpretation of resistance data; and thus, there is a need for the WHO reference library to be continuously updated.
8. **There are different feasibility concerns for the different tNGS platforms**: The overall sentiment that is that all three the tNGS platforms needed to be further developed before it is fully ready for operational use, some more than others. The high level of technical complexity of the sample preparation stages (mainly the library preparation stage) was considered a key challenge for the Deeplex platform, and the need for improved computer analysis and storage capacity was a challenge for the Oxford Nanopore (ONP) platform, though both required a high level of precision and attention to detail, and more steps for early error recognition. The third platform was not ready for testing in two sites. Participants did not want to express explicit preference for one tNGS platform over the other, noting that both Deeplex and ONP had their pros and cons, and that both needed further development to be fit for purpose.

Values, preferences, and equity

The overall sentiment is that that **MDR-TB diagnostic technology needs to balance accuracy, speed, affordability, equity, and cost effectiveness,** and that tNGS technology would need to address these considerations before it can be implemented in these LMIC settings. These values, preferences and equity considerations were consistent across the different stakeholder groups who participated in the study.

1. **Centralized vs decentralized placement may have equity implications for access:** Given the high-level specialised laboratory infrastructure, specialized human resources and technical complexity, tNGS technology it is only suitable for placement at centralized, reference laboratories. This may have equity access considerations as it may mean less access for some regions of the country without reference labs. The MDR-TB case burden of the country is another factor that could influence equitable access at centralized levels. In some settings with high caseloads, the tNGS technology in central laboratories may not be sufficient for processing large caseloads in good time, and in settings with low caseloads, waiting for sufficient samples to batch will also cause delays.
2. **Affordability and cost-effectiveness are major concerns:** There was a major concern about financial costs of the tNGS technology and the affordability for LMICs. Participants were worried about not only the cost of the equipment, but also the costs of ongoing specialist supplies, especially for reagents, as well as the cost of maintaining equipment. They noted that costing calculations should be comprehensive and should include the cost of specialist

consumables, extra general laboratory consumables, and the additional infrastructure needs (such as the extra space, temperature control, and internet connectivity needs). Cost-effectiveness calculations should also be comprehensive, and should include assessment of the impact of the use of tNGS testing on improving TB outcomes.

3. **Synergies should be found to optimize the use of tNGS in diagnostic algorithms:** There was some concern that more and more rapid tests are adopted, without consideration for how to develop synergies between the use of various tests, to optimize use of the new testing technology. Rational use of the new tNGS technology should include reviewing current MDR-TB diagnostic algorithms to determine how and where tNGS technologies would fit best, and what potential synergies could be found.
4. **Strengthening the broader laboratory and health system is needed to optimize the gains from rapid testing:** Optimizing the gains from rapid diagnostics would likely require strengthening the functioning of laboratory and health systems. This includes improved referral systems for clinical samples, closing the gap between clinicians and lab personnel in terms of communication, and training for clinicians to understand and use the test results for effective treatment of MDR-TB. Patient empowerment and education is also needed to improve health seeking behaviour and adherence to treatment.

DISCUSSION

Findings from this qualitative study highlight the high level of acceptability of the value and potential of tNGS technology amongst stakeholders. Nevertheless, there are several feasibility challenges and a variety of values, preferences, and equity considerations to address if tNGS technology is to become operational in routine settings. While start-up problems with the new technology would likely resolve over time, other feasibility challenges would likely remain unless the technology is further refined. For example, the lack of local, in-country supply chains would mean that challenges to specialized supplies and specialized technical support could remain, and affordability and cost-effectiveness considerations would likely require ongoing attention. The implementation considerations raised in this study are echoed in the literature on implementation challenges for new interventions like rapid TB diagnostics. The report outlines a set of implementation considerations for improving the acceptability, feasibility, and equity of future iterations of tNGS technology.

CONCLUSION

Health worker experience and perceptions can influence the uptake and scale-ability of implementing new technologies like tNGS. While there is high acceptability of the value and potential of tNGS technology for MDR-TB diagnosis and management, there are a range of factors shaping acceptability and feasibility that would need to be considered for present uptake and future development of the tNGS technology. Guidelines on the use of tNGS technology, especially in LMICs, would need to take account of not only the effectiveness of the technology in diagnosing MDR-TB, but also the influences of the ease of use of the technology, as well as the organizational contexts and needs, and the values, preferences, and aspirations of stakeholders.

Background

Tuberculosis (TB) remains a major global public health threat. One of the most challenging forms of the disease is multidrug resistant TB (MDR-TB), due to higher morbidity and mortality, complexity of treatment and higher cost (1). WHO estimated that in 2016, close to half a million people were diagnosed with MDR-TB (2). It is estimated that only two out of every three patients with MDR-TB are diagnosed, of those diagnosed, only three out of four are treated, and only half of those treated are cured. This results in the majority (75%) of patients with MDR-TB not being cured; thus, persisting with their illness, spreading the disease and/or dying from their illness. Reasons for this loss of engagement in care include health service factors such as diagnostic delay, treatment delay, inaccurate treatment and patient delays in seeking health care (3).
Timely drug resistance detection is essential to global tuberculosis management (4). The End TB Strategy of the World Health Organization calls for the early diagnosis of TB and universal drug-susceptibility testing (4). Traditional drug susceptibility testing uses culture-based tests which can take several weeks to yield a result and may not be widely available for testing for newer and repurposed drugs (1). Novel molecular testing tools are being developed to rapidly detect TB and resistance to anti-TB drugs, and to diagnose resistance to multiple different TB drugs simultaneously. However, there are obstacles to its wide-scale uptake and implementation, especially in low-and middle-income (LMIC) countries where these diagnostics may be most needed. Obstacles relate to a range of logistical, organizational, and financial factors and there is a need to better understand these barriers. The experiences and perceptions of health workers implementing rapid TB diagnostics can provide useful information to better understand the implementation factors shaping the use and impact of rapid diagnostics.

Genotypic testing, such as Whole genome sequencing (WGS) is molecular-based gene sequencing technology for use in screening and diagnosis of genetic disorders, cancers, as well as in diagnosis of drug resistance in infectious diseases such as HIV, malaria, and TB. Targeted next generation sequencing is a further advancement on WGS, where the focus is on a targeted section of the genome (rather than the whole genome). tNGS for MDR-TB diagnostics is molecular technology that is focused on identifying resistance to TB drugs by targeting investigation of the section of the genome that is known to be associated with mutations associated with resistance to TB drugs. Use of tNGS in TB diagnostics therefore has the capability to provide more comprehensive diagnosis of resistance to a wide spectrum of TB drugs, as compared to current rapid diagnostics that identify resistances for a smaller number of TB drugs. tNGS technology can identify comprehensive multidrug resistance susceptibility in a shorter time (estimated at between 3 and 5 days), and on a sputum sample, as compared to the standard culture-based phenotypic diagnostic testing that takes several weeks (3 to 6 weeks). tNGS technology also provides testing at higher volumes which makes it suitable for use in high TB burden settings. tNGS platforms are aimed at providing an integrated end-to-end solution for MDR-TB drug resistance testing, including automated data analysis and reporting. Diagnostic reports can provide drug resistance information for each sample that was sequenced in a batch, that details which drugs the patient is resistant to (or not). Some platforms are also able to provide information on the lineage of mutations, which can be used to identifying and track strains of drug resistance. tNGS technology therefore offers the potential to provide a comprehensive diagnosis of multi-drug resistance TB. tNGS diagnostic test results would enable timely clinical decisions on appropriate treatment for MDR-TB, which can improve TB treatment outcomes and stop the transmission of resistant strains of TB (5).

Different tNGS platforms have different laboratory workflows, but the general tNGS workflow includes: preparation of the TB sputum sample, DNA extraction from the sputum sample and DNA purification, library preparation, sequencing and data analysis, as shown in the workflow diagram in Figure 1 (5). DNA extraction procedures require specialized laboratory safety levels (biosafety level 3) and special safety equipment and work procedures for contamination containment. Following DNA extraction, DNA library is prepared that involved DNA fragmentation and enrichment, a complex process where it is critical for technicians to exactly follow the instructions of the manufacturers, in terms of use of reagents, controls of temperature and time, and with quality and quantity check required before and after the library preparation. The prepared sample is then run on a sequencing machine and there are

different commercially available models. Computational resources are needed for data analysis, report generation and data storage (5).

Figure 1: tNGS workflow

Targeted tNGS therefore hold the potential, especially in high-burden TB settings, for effective treatment and control of the spread of MDR-TB, but there are obstacles to the uptake of novel, rapid technologies. Obstacles to similar rapid diagnostics have been documented and include concerns about costs, integration into existing laboratory workflows, technical training and skill requirements for utilization of the technology, and the need for expert guidance regarding the management and clinical interpretation of sequencing data (6).

A recent overview of the 'Implementability' of health care interventions concluded that acceptability, fidelity, and feasibility of interventions may influence uptake and scalability and suggested that these factors be considered at the early stages of intervention development and during implementation and evaluation of interventions (7). Stakeholder views of new rapid TB diagnostics are important for informing implementation plans and for understanding the role of stakeholders in shaping implementation and outcomes (8). Health workers play an important role in shaping implementation and outcomes and may be influenced by their experience and perceptions of individual, organizational and system level factors (9, 10, 11, 12, 13). Some argue that health workers ultimately determine how interventions are implemented, based on their understanding of their task, and shaped by their discretionary power in delivering the task (14). For instance, health workers as frontline implementers in low resource settings may be struggling with issues such as chronic staff shortages, multiple demands, and poor performance management, while in high income settings, the high levels of specialization and financial disincentives may shape engagement (9).

It would therefore be useful to better understand how new rapid diagnostic technologies like tNGS are experienced and perceived by those who are implementing the technology. It would be valuable to understand the experiences and perceptions of laboratory staff who are frontline implementers testing out rapid TB diagnostics, for insights on implementation considerations - such as the feasibility, acceptability, equity, and value preferences of new technology (8, 15). The view of TB patients and patient advocates on rapid TB diagnostics can also provide valuable insights, and this would be important to investigate once the new technology is taken up as part of standard care (8).
In preparation for the WHO guideline development meeting on tNGS technology for diagnosis of MDR-TB (5), there is a need to summarize the current evidence on the implementation considerations for use of tNGS. Given the newness of the use of tNGS for TB care, there is little to no qualitative evidence on implementation of tNGS. Hence there is a need to generate primary evidence on experiences and perceptions of health workers who are implementing tNGS.

The FIND Trial

The FIND trial is a multi-center clinical evaluation of the diagnostic accuracy and technical performance of End-to-end tNGS solutions for DR-TB diagnosis in three countries: India, Georgia, and South Africa. The FIND trial was aimed at testing the accuracy of three tNGS platforms in an operational setting of three tNGS platforms namely, Genoscreen Deeplex, Oxford Nanopore (ONP) and ABL Diagnostics. This presented an opportunity to generate primary qualitative data on the experiences and perceptions of implementers in these three settings.

Objectives

This study provides evidence to inform the formation of a new WHO guideline for use of tNGS technology for MDR-TB diagnosis. The objective is to explore the perceptions and experiences of implementers of tNGS technology, with respect to acceptability, feasibility, values and preferences and equity.

Methods

Study design
This was a cross-sectional observational qualitative study using semi-structured interviews.

Setting
The main stakeholders were implementers of the FIND trial in the main reference laboratories that implemented the tNGS trial, located in Georgia, India, and South Africa.

Participants
We were interested to learn about how the experiences and perceptions of the stakeholders who were involved with implementing the FIND trial in the three countries. Participant selection was therefore purposive and representative; purposive as the aim was to include in-country participants with firsthand implementation experience in the three countries where FIND trial was implemented; and representative, because we planned to interview most of the staff that were involved with hands-on implementation. We identified the in-country laboratory staff and management who were directly involved with the implementation and oversight and invited them to participate in interviews to share their experience. We invited staff from the FIND trial, those in coordinating leadership roles and in country project leaders, to provide background information, to share their experiences of running the trial, and their reflections on implementation considerations. We identified three more sites globally where tNGS platforms were being tested in LMIC settings. We invited the researchers from the three sites to participate and all agreed, but one withdrew for logistical reasons. Through a combination of purposive and convenience sampling, we also selected three global experts on TB care and TB diagnostics, to provide a broader view that can further enrich our understanding of implementation considerations of new rapid TB diagnostics such as the tNGS technology. Selection of the global experts was done through convenience sampling, by approaching members on a WHO expert working group on laboratory diagnostics.

We included stakeholders who either had direct experience of tNGS technology or who had insights related to rapid TB diagnostics. The FIND trial did not include use of the tNGS diagnostic reports in clinical decision- making, as the accuracy of the technology is still being evaluated. The diagnostic test results were therefore not yet in use for patient care in these trial settings. We therefore did not include frontline clinicians or patients and patient advocates. It would be important to include clinicians and patients as end users in future studies of the uptake and roll-out of tNGS technology.

Data collection and analysis
A semi- structured interview schedule for implementers was developed with input from the WHO secretariat of the project and FIND researchers. See Appendix 1 for a copy of the interview schedule that was used adapted for different stakeholder groups. The key areas to explore were: How feasible are tNGS solutions to implement? Are tNGS solutions acceptable to end users and other key stakeholders? What are the end-user and key-stakeholder values and preferences related to the use of tNGS solutions? What is the potential impact of tNGS solutions on equity? Is there important uncertainty about or variability in how much people value the main outcomes?

Interviews were conducted in September 2022 through October 2022. All interviews were conducted in English, by the lead author (NL), via a zoom link. Interviews were digitally recorded with agreement of interviewees. The interviewer took notes and wrote a detailed interview summary within 24-36 hours of the interview, that recorded the key ideas and themes from the interview on the key implementation considerations. Where there were gaps or more details required in the summary, this was filled out with information from the digital recording. These quotes were drawn directly from the digital recordings and transcribed verbatim.

Interview summaries were used as the main data source for analysis. As an additional quality assurance step, the interviewer checked if the interviewees wanted to see the interview summary and

check it for accuracy. Most welcomed this opportunity and provided minor edits to improve the summary. The edited versions were used as the final version of the interview summary.

The interviewer read and reread the summaries, to become familiar with the overall data set. Using a thematic analysis approach, key issues were identified within and across interview data sets, and this was synthesized into themes and sub-themes. Key implementation areas identified related to views on acceptability and feasibility, including challenges with setting up and running the technology, ideas about how to improve the technology and implementation, and aspirational views on what future rapid MDR-TB diagnostics should aspire to. The themes and sub-themes were categorized under the 3 main areas of interest in the report: Acceptability, feasibility and values, preferences, and equity considerations. FIND trial implementers had signed non-disclosure agreements that requires confidentiality about their findings. When reporting on the user experiences of the tNGS technology, we will provide general descriptions and only name specific platforms where relevant. For anonymity reasons, we do not name the countries when reporting sentiments and quotes, but refer to the countries as Country A, B and C.

Ethical approval

Institutional ethical approval was obtained for in-country implementers as part of the FIND trial ethical clearance procedures from WHO (WHO ERC # ERC.0003342). Ethical oversight of this work has been reviewed by the Secretariat and is sufficient for the information being collected. All interview participants provided a signed written consent form to participate. See Appendix 2: Information and consent form.

Findings

Demographics

Fifteen (15) interviews were conducted with 17 individuals (in two interviews consisted of two participants each). Participants were from four different stakeholder groups, as shown in Table 1 below: Four stakeholder groups consisted of the main group of tNGS trial implementers in 3 countries (N=8), additional stakeholders with experience of implementation (DIAMA implementers in 2 African country settings (N=2), members of the FIND trial research team (both overall (N=2) and in-country researcher leaders (N=2) who provided background and an overview of experiences from their perspective, and finally, global experts (N=3), who provided a broader expert opinions that help to contextualize the findings further.

Table 1: Description of stakeholders interviewed

Stakeholders	Number interviewed
In country implementers of tNGS in trial setting (Country A, B and C)	8 participants (in 6 interviews- 2 of the interviews had two participants each)
Other tNGS implementers (African consortium)	2 participants
FIND trial research team: leadership management and in-country coordinators (FIND trialists)	4 participants
Global experts on TB and TB diagnostics (Global experts)	3 participants

Acceptability of tNGS technology

There was an overwhelmingly positive sentiment about the acceptability and potential utility of tNGS technology. This was mainly because the technology was seen as major advancement in molecularly MDR TB diagnostics. The main reasons were the comprehensiveness (resistance diagnosis for more drugs and for newest and repurposed drugs), the convenience of using sputum sample (as compared

to culture samples), and the rapidness (quick results compared to phenotypic testing times; 3-5 days as compared to 4-6 weeks). There was also the sense that the technology is arriving at the right time given that resistance to newer TB drugs is likely to increase as use if these drugs become routine.

The tNGS technology represents major advancement in TB diagnostics. The view from all participants was that tNGS technology represents a major advancement in TB diagnostics. The ability of the technology to provide a more comprehensive diagnosis of TB drug resistance was considered a major positive impact:

> *"What would be the added advantages of having tNGS in our environment? It would be massive! It will give us more comprehensive information on resistance type…We'll have more drugs than we are currently seeing and what we're getting in our diagnostic services. You could probably get everything… you could get everything you are interested in" (Country A, participant 4)*

Participants were excited about future development of tNGS technology. Some described it as a 'revolutionary' step forward in MDR-TB diagnostics, while they also noted that the technology is still evolving:

> *"It's really a revolutionary stage in diagnostics. And as [participant 2] mentioned, having this kind of opportunity which can easily replace all the current diagnostics tests, … that will be quite effective….We've never been so close to the best centralized diagnostics in TB. There's always room for improvement. But I'm sure all these companies will be working on some improvements, and we will see big developments in the future. I'm sure about that. (Country B, participant 1).*

The participant cautioned that there is still some way to go before one would replace the current phenotypic testing for drug resistance for MDR- TB:

> *"The transition process itself is not that simple, because the culture is still gold standard, even for WHO. So, we have to have the culture anyway. So, moving from phenotypic to totally genotypic diagnostics is a huge leap for TB diagnostics, for sure." (Country B, participant 1).*

These positive sentiments were echoed also by implementers from country C, who highlighted the importance of faster drug susceptibility testing for MDR-TB:

> *"TB diagnostics have arrived, really. Its diagnostic utopia where TB is concerned. With other bacteria, were struggling, we just don't know… I mean here we have the whole genome in just two days! How cool is that! … Unbelievable, really." (Country C, participant 1).*
>
> And further:
>
> *"It's really wonderful to be in this space, you know, that we are able to do stuff that will be so important in the future… I believe these technologies must be out there because we have to give answers fast, in a more, better structured way to our patients. I think we owe them [patients] that. And we've had enough of taking months to give a drug susceptibility test in the phenotypic way…. The fact that we are here is great. Its wonderful technology, we believe in it. Please get it to countries, we need it." (Country C, participant 1).*

The reasons for this positive view included the fact that this molecular tNGS technology can be used directly on using a sputum sample, as opposed to phenotypic testing that require a culture sample. The second reason is that the technology can provide comprehensive information on drug resistance on detection many TB drugs (more information on up to 14 drugs), both for first and 2nd line treatments. And the third reason is the short turnaround time of receiving test results of 3 to 5 days from starting the diagnostic testing process, as opposed to 4 to 6 weeks for phenotypic testing.

The short turnaround times for receiving test results was considered a major benefit, especially in the light of shorter TB treatment regimens for multi-drug resistant TB. With a treatment regime of 6 months, the concern with phenotypic testing is that results for drug resistance is received only 2 to 3 months later, when the patient is already a third of way or halfway through the treatment regime.

There is a good window of opportunity for utility of tNGS. The sense is that drug resistance to 2nd line TB drugs is still low, but likely to rise in the next 5 years. There will be a growing need for this technology in the next few years and some thought this was a good window of opportunity to build on this technology:

> *"It depends on what algorithm that the country is following. [Country C] is doing good, we are up to 3 days for [diagnosing] pre-XDR, in 3 days we know if its pre-XDR. … And currently the rates of resistance to Bedaquiline, Pretomanid and Linezolid are low. We are getting away without doing*

the test because the resistance is low. But in another 4-5 years we still need this technology. So, it's time to build on this technology, now, so that we are experts at it when it happens, when resistances start coming." (Country C, participant 1).

Feasibility considerations of tNGS technology

The overall sentiment was that the tNGS technology needs to be further developed before it can be considered fully ready for operational use. Implementers experienced a range of challenges with implementing the tNGS technology which influenced their views on the feasibility of the technology in LMIC settings. The challenges they identified include the starting up problems, the high level of specialization of human resources and of laboratory infrastructure required, the supply chain challenges, and the high level of technical complexity required for some of the platforms. The technical challenges related to the high level of complexity of the library preparation stages, and problems with data management and storage, and concerns about data security. Continuous updating of the WHO mutations reference library would also be required for the tNGS technology to be useful, as the reference library was a central tool that allowed for meaningful interpretation of test results. Although there were different types and levels of concerns for the two main platforms tested (Deeplex and Nanopore), the general sense was that the tNGS technology has still some way to go to be fully ready for operational use.

Start-up challenges

The trial sites in all three countries were high level, centralized laboratory settings, where the focus was on both researching diagnostics and providing diagnostic services. This meant all three implementation sites had specialised laboratory infrastructure and specialised human resources to implement molecular diagnostic testing.

Procuring and setting of challenges: Setting up processes were logistically complicated. Given the newness of the tNGS technology, implementers felt they were not always clear about what specialist equipment and supplies would be needed, and how to procure those. So, they were not able to anticipate and prepare for the delay they experienced. For instance, multiple specialist components, small and large had to be imported from European countries, and complex importation processes caused procurement blockages. One consignment of equipment got lost, and another got stuck at customs for 2 months. In the African consortium countries, procurement of equipment and supplies made it difficult to fully implement the tNGS platform.

Preparation and training requirements: The intensive one week of training provided for two of the platforms worked well. Training for the third platform was incomplete and resulted in this platform not being implemented in one country, and implementation was delayed in another country. As a result of set-up delays, there was a gap of one to two months between the training and the testing phase of the tNGS technology, which was not ideal. Training involved observation of one trial run of the full tNGS procedure which was useful. Implementers noted that given the complexity of the procedures, they would have benefited from doing another hands-on run through, under supervision of the trainers. There were several problems with initial failed runs which required the involvement of the manufacturers to resolve problems. There was a lack of local technical support from local vendors as the technology is so new and, in some places, there were no local vendors. In the African consortium countries, the lack of technical support with sorting out a faulty machine, resulted in a two-year delay in implementation in one country, and in the other, implementation of the tNGS platform was put on hold after the completion of the testing phase. Implementers recommended that manufacturers make available a checklist of what errors to expect, and what the trouble shooting steps are.

Specialized human resources and infrastructure requirements

Specialised human resources are limited to central laboratories: As tNGS is a molecular-based testing platform, the platform requires specialist laboratory scientists, such as molecular or medical scientists. In all three countries, molecular scientists were in place, and were leading a laboratory team that

included laboratory technicians. The scientists had experience in molecular diagnostics, were highly trained (some had PHDs), and some had advance experience in testing other rapid molecular TB diagnostics. Molecular scientists were only available at reference/central laboratories, and not at regional laboratories, which means that in these settings this test could not be decentralized to lower-level laboratories.

The tNGS technology was thought to be technically challenging even for specialist staff with molecular testing experience, making it difficult to consider the use of tNGS in the absence of specialized personnel. It requires specialist staff *"who are not just trained, but who has acumen to spot errors*" (Country C, participant 3). An implementer explains how the need for specialist staff limits the technology to being placed in a centralized laboratory:

> *"If we take it down to a routine environment where we're going to have technicians and technologist run this technology, it won't work. So, in terms of HR, a key component as well, that you need someone skilled, in its current format, to perform this test. If we move away from this intensive, technically challenging preparation to a more automated format, which is really possible, that would improve [the test], or adoption would be much easier." (Country A, participant 4).*

There was also concern in some settings that specialist personnel such as molecular scientist are difficult to recruit, and that retention of specialist staff was becoming a challenge due to out-migration of skills. It was noted that in some LMIC settings, there is a need to strengthen policies for not only training and recruitment of such specialist personnel, but also for adequate renumeration and retention.

Specialized laboratory infrastructure is mainly available at central laboratories: Molecular testing needs specialized infrastructure with different rooms for different processes (e.g., 'wet' and 'dry' rooms), designed for containment of contamination. These room and the specialist cold storage facilities were available at most central laboratories, but not necessarily at regional laboratories. While specialist infrastructure was in place in the testing laboratories in the trial sites, some required extra resources and adaptations. For example, in Country C, they had in place the 4 different rooms that required for their routine molecular laboratories but found that the protocol for one of the tNGS platforms required a 5th separate room. This was not anticipated and meant the team had to scramble to find additional space. In this same laboratory, they also had to add an additional air conditioner and a boosting air conditioner unit, another humidifier, and a special table for one of the instruments. The team felt that it would have been helpful to be aware of these additional requirement earlier in the process. A participant points to the need for ensure the infrastructure is in place when considering use of tNGS technology.

> *"Our model is designed to support routine work… that is not technically challenging. TNGS is a really complex methodology. The infrastructure is not readily available…A simple thing like a degree fridge, is not readily accessible across the bigger labs in the country. They require minus 70 degrees storage, mines 20 degrees storage. Those things are not readily available… Or if they are available, do not have sufficient space to keep these type of products immediately… Before the technology can be introduced, we need to ensure that the infrastructure for the technology is supported." (Country A, participant 4).*

The platform requires uninterrupted supply of electricity: The process from preparing the sample for sequencing to producing a diagnostic report with the TNGS platforms takes between 3 and 5 days. This involves overnight runs, so an uninterrupted power supply is required for the full duration of the process. In country A, interruptions in power supply is a regular occurrence, and even though backups are in place, this is not sufficient to prevent disruptions to the testing process. Luckily for this team, there were no major power cuts during the period of the FIND trial. Staff noted that that it was fortunate for them. Interrupted power supply also meant staff had to do additional monitoring to ensure cold storage conditions for specialist kits were being maintained and transferring stock to optimal storage when required.

The platform requires high-capacity computer systems and internet connectivity: All three sequencing platforms rely on access to computer systems, and some rely on access to web-based platforms for data analysis and generating of reports. The data analysis component of the sequencing

platforms generate large data files which rely on high capacity, stable internet connectivity.

> *"To implement this technology, computing power, storage infrastructure and internet connectivity are extremely critical…The two platforms we tested are reliant on internet connectivity. And we're talking about uploading massive sizes of files of a server for analysis. If a country doesn't have the infrastructure to support internet and uploading info into a cloud for analysis, you can forget it. I don't know if they are considering having a localized machine to do that analysis.) (Country A, participant 4).*

Supply chain challenges

Procurement of tNGS supplies encountered disruptions and delays: A reliable and continuous supply chain for stocks is needed to operate the tNGS platforms. Disruption in the supplies was noted as a major challenge across the three trial sites, as well as in the African consortium sites. The NGS platforms all required specialist kits, containing, specialist reagents and laboratory consumables that could not be procured locally. These specialist supplies had to be ordered from overseas, which took time and caused delays. Initially due to the newness of the products, it was not always clear what supplies would be needed, by when, and how to procure it. Bottle necks in the procurement process caused delays of three weeks or more in receiving orders. Procurement challenges were worse in the African consortium sites, resulting in delays of up to 6-months. Multiple delays due to supply chain blockages jeopardized the implementation of the platforms and caused staff much anxiety: As one implementer noted:

> *"We cannot be running after reagents… Our stress is not doing the job, our stress is the logistics." (Country C, P1)*

The lack of local vendors for specialist supplies and for maintenance and technical support also contributed to procurement challenges. Even where local vendors were available, they were not able to speed up procurement or provide the technical support needed. These challenges resulted in only partial implementation of the tNGS platform in the African consortium.

Specialist products had a short shelf-life of 4-6 months which meant products had to be used quickly before it reached its expiry date. This required careful planning and stock management which put further pressure on staff to ensure supplies were used in time. In country C, the staff had to scramble to rush to ensure the products were used before it expired. Another concern was that the specialist kits had specific storages and other requirements during transit, and there was the danger that supply chain blockages could result in the supplies being ruined. In country (C), implementers feared that their consignment may get stuck on airport tarmac for hours, in high temperatures, which would have ruined it.

A different procurement challenge was the larger volumes of general laboratory consumables that were required alongside the specialized kits. Some platforms needed this more than others. Implementers in country A reported that for one of the platforms, they used four to five times what they used for other routine tests. In country C, implementers reported that they not only needed large amounts of general laboratory consumables, but that some of those consumables were "very particular" and not in routine use. It took some scrambling to get these consumables and staff had to make their own adjustments to fit the specific requirements of the tNGS platform.

Technical challenges with implementation of the sequencing platform

The high level of technical complexity of tNGS technology was a challenge: tNGS technology is high complexity molecular testing and the sense is that it is technically challenging even for highly trained and experienced molecular scientists. Preparing the sample for sequencing involves multiple steps, from cleaning the sample to library preparation, sequencing, and data analysis.

The sense was that there are too many steps and too many steps that require a high level of precision with measurement, and that this leaves too much room for error. The complexity comes in terms of the large number of steps and the level of precision required for many of the steps, the very small measurements involved, the careful transcription of information, and the critical issue of avoiding cross contamination of samples. One participant commenting on the sample preparation for the Deeplex platform said it was very challenging to work with "*very small amounts of reagents to ensure*

the correct volume is added", which is *"demanding of precision*" and requires *"the utmost concentration"*, and with" no *room for error"* (Country A, participant 1).
Some of the problems implementers experiences were contamination of control samples, recognition, struggles with uploading data from the sequencer to a web-based data analysis tool, problems with data analysis software, problem with computer storage capacity, problems with sensitivity of equipment to temperature and disturbances. In one setting there was multiple failed runs where the source of the problem proved difficult to pin down even with the help of local technicians- and it was eventually related to seasonal changes in quality of water supply to the laboratory.
To increase its acceptability and feasibility, the library preparation steps need to be simplified to reduce the complexity (through for instance consolidation and automation of steps). Simplification will allow less specialised laboratory staff to implement it.

> *"In summary, it's almost there. There's certain areas that need polishing. But for me, the biggest concern is the library preparation. It's not diagnostic ready, definitely not diagnostic ready. … There's too many steps, 30 steps, 40 steps. I mean, really, with diagnostics, you don't want 40 steps." (Country A, Participant 4)*

The concern is that these complex sample preparation steps (described as 'laborious' in Country C), becomes even more complex, with more room for error when applied to the preparation of multiple samples for a batch (up to 48 samples for one of the platforms). A related concern is that there are insufficient opportunities for timely detection and correction of error in the sequencing and analysis process. For instance, for platform A, an error in the sequencing phase can only be spotted 2 to 3 hours into the run. Also, once the cartridge is loaded it cannot be reused, which is a source of waste.

Data analysis and storage presented challenges: For all the platforms, the data analysis steps were automated, thought the degree of automation differed across platforms. For two of the platforms, this involved uploading to a web-based analysis tool for two of the platforms. For another platform, a computer was provided with customized analysis software. The data analysis processes were not always smooth. For example, in country A, the sequencing machine for the Deeplex platform was not linked to a computer, so this required manual uploading of data files to the web-based analysis tool. While this did not require specialist skills, the process required careful transcription and making sure the files are labeled correctly. Implementers thought this was a laborious analysis process. It also left too much room for mistakes during the transcribing and uploading steps. For one of the platforms, there were problems with initial uploading of data files that required help from the manufacturer. For another platform, initial problems due to bugs in the software also needed intervention from the manufacturer. A participant explained how the lack of an automated links between the raw data and the analysis tool is problematic:

> *"With [platform B], the software was challenging. There were a few bugs with the analysis. But the analysis tool is not well designed. There is too much room for error, for mistakes. There's no link between the raw data and the analysis tool. It's an independent step. So, you take the raw data, and you tell the analysis tool what each file each, which sample each file is. That's dangerous. The raw data should already inform the analysis tool what is the raw data that its reading. Because there's too much room for error there." (Country A, participant 4)*

Data analysis generates very large data files, and this posed problems in some cases. In one setting, where they performed extra tests, the computer provided for data analysis ran out of storage space, due to the large size of the data set. With the help of the manufacturer and their local IT specialists, they had to try different solutions before solving the problem. In the end they used additional external hard drive storage capacity. However, while fixing the problem, data was lost and could not be retrieved, and they were required to repeat some of the runs. They expressed concern that there should not only be sufficient storage space, but that there should be adequate systems for backing up data to prevent data loss. A related challenge associated with the analysis of large data files, is that it takes time. It usually takes seven to twelve hours of overnight runs. Successful runs require stable, high speed internet connectivity which is not always available.

Utility of diagnostic reports and informational support from the WHO mutations catalogue: The evaluation of the tNGS platforms in the trial sites stopped at the stage where a diagnostic report with the test results was produced. The trial did not involve the evaluation of clinicians' use of the test result report for clinical decision-making. Participants nevertheless shared their views on the potential

the utility of the tNGS reports for clinical decision-making.
The process of producing a diagnostic report with test results is automated for all the three platforms. Reports differed in the level of detail they provided, and reports were presented in a format that could be used for clinical decision-making. For each drug, information on resistance is expressed in terms of whether there is resistance or susceptibility to the drug, and the confidence level of that test result (high, medium, or low). Identification of mutations are based on the WHO reference catalogue of mutations. Data from platform A could also be used to detect mixed infections. A second set of information is information of lineage of mutations, which allows for surveillance of transmission trends that can be useful for infection control and outbreak management.
Some participants pointed out that there is a lot of diagnostic complexity involved with diagnosing resistance to newer drugs like Bedaquiline. There are still areas of uncertainty in our knowledge about the drug resistance status of new and repurposed drugs and more research evidence is needed. All stakeholders noted that the WHO mutations catalogue is the most important informational support tool for the tNGS platform, and without this, the diagnostic test would not produce meaningful and useful test results. Further development of the tNGS platforms is dependent on the expansion of our knowledge about mutation and that for this, the annual updates to the WHO mutations catalogue is critical.

> *"In terms of maturity- the reference library is key…As an example, Bedaquiline, it's really new. We don't have a comprehensive list of mutations that are associated with resistance, or not associated with resistance. The WHO is working on it. But there is a lot of work that need to be done, to actually create this complete list mutation catalogue, where we can then decide, its susceptible, or resistant… It doesn't only apply to Bedaquiline, it applies to a lot of other drugs, especially new and repurposed drugs." (Country A, Participant 4).*

Data security mechanisms and interlinkages of information systems are needed: Implementers were asked if they had views or concerns about data security given that there are different analysis software in use. For two of the tNGS platforms, data is uploaded to web-based analysis platforms, and for one the platforms, the data is uploaded to a computer with software that is linked to the platform. Some thought that deidentification of data may provide sufficient data security, at least during the study trial phase. They recommended that in addition to de-identifying the data, mechanisms are needed to regulate what happens to the data. In the African consortium setting, they had a developed a memorandum of understanding with the manufacturers, regarding the data management and security; for example, it allowed for the data to be destroyed after one year for security reasons.

> *"Manufacturers should make this information exchange highly secure the way that it guarantees personal data safety. At this point data belongs to the sponsor, as this was a research project requested by FIND and for data analysis part and conclusions, we all agreed clearly on data ownership and non-disclosure form our side. For the future use the best will be if software will be integrated into the existing local information systems." (Country B, participant 1).*

Others were more concerned about data security, especially the lack of clarity about what happens to the data once uploaded, and the uncertainty about data ownership. In country A, there was concern that the laboratory did not have control of the storage of the data.
Implementers highlighted the need for integration between the tNGS information system and the routine laboratory information system. There needs to be interlinkages and bi-directional communication between the two information systems, in a way that allows for secure and automated two-way communication. One consideration is the need to balance the need for efficient and for secure data transfer between the two systems. In country A for example, the routine information network security requirements results in data upload and transfer that can take hours, and some long uploads fail during these long transfers. The concern is that the secure routine laboratory information network would not easily handle large datafiles associated with TNGS platforms. A participant wondered if a separate network would be needed to facilitate fast, accurate and secure data transfers:

> *"But to integrate it with our system, something that is key, is interfacing it with our lab info system….Sending the information back, it has to integrate into our lab information system, without us needing to enter it….To improve the accuracy about how it is captured, there will have to a bi-directional communication between the software [of the tool and the lab info system]. And I don't think any of these tools have that capability right now." (Country A, participant 4).*

Differences on feasibility views across different tNGS platforms

The overall sentiment that is that all three the tNGS platforms need to be further developed before it is fully ready for operational use. However, there were differences in the sentiments about feasibility challenges, with some being regarded as having more substantial challenges. The high level of technical complexity of the sample preparation stages (mainly the library preparation stage) was considered a key challenge for the Deeplex platform. Sample preparation prior to sequencing was considered less of a challenge for the Oxford Nanopore (ONP) platform. The concern about the need for laboratory professionals to have a high level of specialist skill was therefore more linked to the Deeplex platform, but both nevertheless require a high level of precision and attention to detail. For both the Deeplex and ONP platforms, there was concern that the technical requirements and process flow of the tests does not allow for sufficient and early opportunities for error recognition and error correction, which could potentially result in wasted test runs on both platforms. It was suggested that for the Deeplex platform, there needs to be more integrated, built-in electronic linkage between the sequencer unit and the data analysis platform (to allow for a more seamless end-to-end solution). A concern for the ONP platform was that the computer analysis and memory capacity was not able to fully support the volume of testing that was done in one site, and in another, the technology was found to be too sensitive to disturbances and unstable internet connectivity. Participants were not able to comment on the 3rd tNGS platform as it could not be implemented in 2 sites due to start-up and training problems, and in one site, it was too early to tell as implementation had just started up. Participants did not want to express explicit preference for one tNGS platform over the other, noting that both Deeplex and ONP had their pros and cons, and that both needed further development to be fit for purpose. Participants noted that while the start-up problems with the new technology were understandable and likely to be more easily resolved over time, some of the feasibility challenges would likely remain unless the technology is further refined. For example, given the lack of local, in-country supply chains for specialized supplies and specialized technical support, these issues would likely remain a challenge. Affordability and cost-effectiveness considerations would likely require ongoing attention.

Values, preferences, and equity considerations

The overall sentiment is that that MDR-TB diagnostic technology needs to balance accuracy, speed, affordability, equity, and cost effectiveness, and that tNGS technology would need to address these considerations before it can be implemented in these LMIC settings. These considerations are detailed below. These values, preferences and equity considerations showed little variation across the different stakeholder groups who participated in the study.

tNGS technology needs further development for improved acceptability and feasibility

Further development of the tNGS technology is needed to improve its field readiness: While stakeholders had high praise for the diagnostic advances represented by tNGS technology, some also had reservations about whether the technology was ready to be implemented in operational settings. The reservations were based on their experience of logistical challenges described earlier, such as with setting up, the high technical complexity test and challenges with running the tests. While some of the tNGS platforms were considered more ready than others, one was thought to be still at an early stage as it was ready to be implemented in two countries, and was much delayed in the third country. Also, the sense was that the technology was not yet a full 'end-to-end solution' as it relies on different components from different manufacturers, and/or the components are not fully integrated:

> *"Further, these technologies are dependent on several ancillary pieces of equipment… It's not an all-in-one package…It's not just that you can buy the kit and you run the test. You need a DNA quantification tool. You need homogenizers…. magnetic plates. These are all things that are not supplied with the test. So, in terms of infancy, like I said, the technology needs to mature…. [From] my experience, the companies commercializing the products, they are not market ready. There is a lot of work to be done." (Country A, participant 4)."*

Areas of concern that would need to be addressed in future development of the tNGS platform include the need to reduce the level of complexity of the sequencing procedures and develop more automated interlinkage with the data analysis processes. Mechanisms would need to be in place to ensure a continuous supply of specialist products, especially the reagents. Local vendors are needed for the maintenance of equipment and to provide technical support.

A lower complexity tNGS platform is needed to decentralize the use of tNGS technology: A key concern was the high level of complexity of the tNGS platform, and the associated specialized human resources and laboratory infrastructure. As the complexity can lead to human error it limits the test to centralized laboratories and thus limits its accessibility and usefulness. The test procedures should be lower complexity to enable routine use at decentralized laboratories. This will require consolidation through reducing some steps and standardizing others, including automation of steps. For instance, semi-automated robotic liquid handlers are one tools that can improve the process flow and reduce the level of human error:

> *"So, if we just automate the library preparation component of the assay, then we'd have more suitable platform or end-to-end solution that can be done routinely at any level, at our [regional] diagnostic environment." (Country C, participant 4)*

Simplification will also require reengineering of laboratory processes to allow for more test automation. It will also require training and capacity building of technicians who routinely work at these laboratories.

> "*Building capacity to me means giving opportunity to technologists who actually do the assay… You want people at the coalface to be able to do the assay, to be trained and understand the data, who can trouble shoot, who can become professionals at tNGS. And I think that is the future." (Global expert 3).*

Ongoing evidence is needed on resistance status of drugs for tNGS technology to be useful: It was noted that we do not have sufficient knowledge on the resistance status of new and repurposed drugs, which will limit the usefulness of tNGS technology. More research is needed on the resistance status of these drugs. The WHO mutations catalogue is a key source of informational support for interpreting the drug resistance results of tNGS platforms. It was noted that for tNGS technology to be useful, we need to have continual updates of the WHO mutations reference library to enable us to interpret the tNGS test results meaningfully:

> *"But the technology, tNGS is only as good as the info we have about mutations that are associated with resistant drugs and genetic targets." (Country A, participant 4)*

More evidence on operational feasibility is needed: Molecular testing requires not only basic laboratory infrastructure**,** but specialized infrastructure. To promote tNGS technology in LMIC settings, some global experts felt that we need to have more pragmatic trials to test it out in many different environments, especially where there are infrastructure limitations.

> *"We need to put these things in laboratories from Cape Town to Cairo and finding out how does it work. Where do things go wrong." (Global expert 3)*

The participant explained the dilemma facing a colleague who wants to implement tNGS in their country, when some of the basic infrastructure is not in place:

> *"A colleague wants to implement tNGS. They have the instrument. Procurement of consumables is not easy. They don't have good internet. They have a fantastic server, but no internet. And they've got power constraints. Can you run a fancy machine in a hot tropical environment?*

Aspirational views on tNGS as a point of care test: Stakeholders noted that effective diagnostic testing for MDR TB needs to be as close as possible to the point of diagnosis and treatment of patients. For this reason, several implementers expressed the wish that future tNGS and related technologies could aim for point of care testing as their end goal.

> *"…this is a brilliant development for the centralized TB diagnostics. But [what] we will still be missing is the point of care tests, and that's very important… To remember that diagnostics happens really close to the patient care." (Country B, participant 1).*

Others also wished for the technology to be sufficiently automated to be placed at the point of care:

"If the technology can become like a 'plug and play', it can be placed closer to the point of care" (Country C, participant 1).

Equitable access

Centralized, reference laboratory placement requires tradeoffs in terms of equitable access*:* Stakeholders agreed given the complexity of tNGS technology (and the specialist laboratory and HR requirements), it is only suitable for placement at centralized, reference laboratories that can perform molecular testing. This has equity considerations as it may mean less access for some regions of the country without reference laboratorys. The MDR-TB case burden of the country is another factor that could influence equitable access at centralized levels. In some settings with low caseloads, waiting for sufficient samples to batch will cause delays in running the samples.

"If we need to do batching for 51 sample matching, then that will take too long. We need to know faster what the treatment is…." (*Country C, participant 1).*

This participant jokingly referred to the affordability aspect of this trade-off between batching capacity and affordability of the tests.

"We need a plug and play, for one sample at a time. If you can give me 48 [sample capacity] and charge me the same rate as for doing one sample, I'm cool with it!" (Country C, participant 1).

In settings with high caseloads, having one centralized machine may not be able to cope with the large number of samples. A participant explains:

"The technology at this time is ready for a reference lab level. However, it wouldn't be able to sustain the volume that is required. Then you'd have to make key decisions on which samples would need to be tested using this technology. Ideally everyone would want all their samples to be tested. We' got about 15- 20 000 samples a year. That is something that is not feasible for one lab to perform. You're looking at 60 samples a day and the turn-around time of the technology is up to 3 days a run, so there no way you can. You'd need to have sufficient capacity to deal with the entire country's burden." (Country A, participant 4).

Affordability and cost-effectiveness

Affordable and sustainable funding models are needed for tNGS in LMICs: Stakeholders across settings expressed a concern about financial costs of the tNGS technology and the affordability for LMICs. They were worried about not only the cost of the equipment, but also the costs of supplies, especially for reagents, as well as the cost of maintaining equipment. Other considerations would be how the use of tNGS technology would be funded, where this would involve a model of donor subsidies to make it affordable, and how sustainable a subsidized funding model would be. One of the considerations is how one gets the economy of scale that can bring tNGS testing cost down. One participant illustrated this dilemma by comparing with the scale of testing done during COVID:

"One of the strongest points is the cost, how does one bring the cost down. With COVID, they threw money at COVID. You get thousands of people with the disease every day. So, you can take your 100, 200, 500 samples and sequence dirt cheaply every individual…Because you put a whole lot of sequencing into one run. But with one individual, the cost of that individual is enormous. So how do we get to that point where the cost is $20, $10…. (Global expert 3)

The participant added:

"I joked with colleagues five years ago, that ideally we should be going for $1 WSG, for a whole genome. That's where we should be going! Because that means that everybody could be implementing this type of technology and get the benefit from this type of technology." (Global expert 3)

Implementers thought that costing the use of the tNGS platforms should be comprehensive. It was felt that even if the equipment was made affordable, that cost of supplies, such as the reagent cost would be substantial. Costing should include the cost of specialist consumables, extra general laboratory consumables, and the extra infrastructure needs (such as the additional space and temperature control measures). Where computer costs are involved, the costing should include the need for a computer with large computational and storage space, the need for back-up storage facilities, and the cost, maintenance and lifetime replacement cost of the computer and software programmes. Costing

should also include equipment maintenance and servicing costs. There may also additional costs associated with securing high internet capacity and interlinking the tNGS information platform and the laboratory's own information system.

Cost effectiveness should include test effectiveness and impact on TB treatment outcomes: While there is appreciation for dramatic improvements in scope and turnaround time of diagnosis of MDR TB, the real value should be measured in terms of the accuracy of the test and the impact on TB treatment outcome. Test results should be accurate and valid, and the probability of uncertainty should be clear so that it gives clinicians directions to follow for how to manage the patient, and this should be balanced against the need for speed in getting the test result. The effectiveness of the test (accuracy and the validity) to diagnose drug resistance should also be considered in cost-effectiveness calculations. This means that the rates of indeterminate results of the tNGS platforms should be part of costing (in terms of the time and cost of repeat testing). There was a senses that delays of a few days should be acceptable to ensure the results are accurate, rather than having uncertain results available in one day.

Another complexity raised by a global expert, was the need to consider doing cost-effectiveness of using the assay directly on sputum as compared to using it on culture samples. This is to take account of the different levels of accuracy the test may have on sputum versus culture samples, for different TB population. The argument is that as one needs a lot of bacteria in sputum to get good test results, there may therefore potentially be the higher levels of indeterminate test results on sputum samples. If so, then this high level of indeterminate results would need to be factored in to cost effectiveness calculations. According to the global expert, one scenario that may need to be costed, is using the tNGS technology in a different ways for different groups. For example, it could be used in an "indirect" way, on a culture sample rather than a sputum sample, where the chances of successful test result (when applied to a culture sample) would presumably be very high. Using such an indirect test approach may be advisable with patient groups with low bacteria sputum, such as in high HIV prevalence areas. The point was that although using the test on a culture sample would cause a delay of about 14 days (time needed to cultivate the culture sample), it may be worth it if the technology produced a higher chance of getting successful test results, and this may therefore be a good way to optimize value for money.

There was a sense that the real value of tNGS is whether its impact is improved MDR TB treatment and population health outcomes.

> *"Looking at a program, when you put a tNGS into a program, we look at how its impacting patients' lives, how many lives are being saved….We need to look at whether its decreasing TB or not. We have to look at how may days to treatment, days to effective treatment…. For example, we take a clinic here, and you look at people who are using tNGS and who are not and have a comparison [of impact]…. So that's the only way we can really know, how good or bad it is." (Country C, participant 1).*

Synergies for optimizing diagnostic algorithms

Need to rationalize and integrate the use of different types of rapid diagnostics: While the value of tNGS is recognized, some stakeholders were concerned that more and more rapid tests are adopted, without consideration for how to rationalize the use of these tests. The sense is that this results in inefficiencies, and that one needs to consider a more systematized way of adopting rapid TB diagnostics:

> *"One of the problems of the diagnostic market is, there are tests coming in and we are grabbing all these tests and grabbing and grabbing. And none of them is substituting another. So, we have smear, we have LPA, we have GeneXpert, we have Culture… So, one thing to think about is really having one test which can substitute others, so we don't have all these different stages, different tests. So, I see this tNGS as the future of diagnostics, but it needs to take over from the other tests. There is no point of this work if were still running all the other tests." (Country B, participant 1).*

Rational adoption of new rapid tests such as tNGS should also include considerations for if and how it would fit with a transition from phenotypic testing towards genotypic testing; how phenotypic testing may still be needed in future scenarios. And costing of the use of tNGS should include the costs of the scenario where both genotypic and phenotypic testing for MDR TB would still be required.

A similar point was made by another participant who pointed to the need to be clear about the purpose of the rapid diagnostic in the overall cascade and the need for it to be easily implemented:

> *"I need a technology, a technology. I'm not saying it should be targeted next generations sequencing. I just want a molecular technology that is not too technical, that is not too demanding in terms of energy and whatever resources, that would give you the necessary information to treat the majority of your patients. Because the smaller proportion of patients that are difficult to treat need a bit more time. You can do phenotype; you can do the whole genome sequencing. But your high throughput, your large numbers need to be sorted out first." (Country A, participant 3).*

Further, there are several other issues to consider when deciding on rational use of tNGS technology. These include the burden of disease of the country, the different entry levels for TB patients for TB diagnostics, where the tNGS platform would fit in the testing algorithm. In considering cost-effectiveness, will tNGS replace other technologies that is superseded, like LPA? And what other tools are available to do the same or a similar job? Will the technology be useful beyond diagnostics, to include targeted surveillance in places for better insights into emergent strains?
Not all national TB programmes may need such complex technology. Countries with no means to afford tNGS technology, could consider placement of the technology in a supra reference laboratory in one country, and then share access to the tNGS technology, as part of multicounty regional collaborative. This is the approach used in some of the African consortium countries. Finally, in LMIC settings supported by external donor funds for control of infectious disease, it will require careful consideration of how to balance the cost of maintaining current programme funding levels against acquiring costly new technologies.

Potential for synergies in using tNGS technology for other diseases: Participants wondered about how the use of the tNGS technology could be optimized, such as for example, using it for different types of MDR-TB diagnostics, as well as for research purposes

> *"…But the additional thing is that we can proceed with not only MDR, but also with mono resistant cases. For example, mono resistance for Rifampicin, or mono resistance for Isoniazid. That will also make the batching system easier for us. But also depending on the purpose of the lab; researching… Could we mix the research samples with the diagnostic samples?... So, it depends on what road you are taking, and whatever questions you have and what are your motives." (Country B, participant 2).*

Or whether the equipment could be used or adapted for other diagnostic uses, such as for HIV drug resistance or for prenatal genetic testing. One participant raised the issue of exploring if tNGS sequencer technology can be adapted from being mono-pathogen to being poly-pathogen to make the technology more economically viable:

> *"One way of solving this problem [of costs]…We want to explore a study to say, why don't we make the sequencer, instead of being mono-pathogen, to make it poly-pathogen. And have that with the technicians with their own specialty, one' looking at Ebola, one's looking at TB, one's looking at Nile fever, etc. That may make it economically viable" (Global expert 2).*

While synergies may be possible, the sense is the focus should be on feasibility for use in TB care:

> *"I don't mind if my machine is going to be used for oncology, HLA etc., I want it to be cheap for my TB patients!" (Country C, participant 1)*

Strengthening the broader laboratory and health system

Health system strengthening to optimize the tNGS diagnostic gains: While there is appreciation for the value of tNGS for comprehensive and quick diagnosis of MDR-TB, some felt the gains from rapid and comprehensive diagnostics can only be optimized if there is also increased efficiencies in the laboratory and health care system more broadly. When asked what was needed to optimize the gains

from rapid diagnostics, one participant explained the importance of "reengineering" improvements in the laboratory and the broader health care system:

> *"A colleague of mine who was involved with the LPA evaluation said that we have to reengineer the lab to do these tests. But if we don't reengineer the health care system, the value of the time that's gained in the lab is lost in the health care system. I think that's really quite important. You can have test that can give you an answer in 2 days…You ship that test result to the clinic. Then the clinic has to find the patient. And the patient doesn't respond, for argument's sake, he's lost. I can't give you the numbers, but I can say it's quite severe. A certain proportion of patients that come in for the test, the test gets done, but they never come back to the clinic…" (Global expert 3)*

Improvement of laboratory and the health services and systems are needed to allow for optimal TB treatment care and patient health outcomes. Areas that may need to improve include the referral systems for clinical samples, closing the gap between clinicians and lab personnel in terms of communication, training for clinicians to understand the test results and to use the test results for effective clinical decisions and effective MDR treatment option. Ensuring an effective supply chain of drugs is key. Patient empowerment and education is also needed to improve health seeking behaviour and adherence to treatment.

The need to make the tNGS reports useful for clinicians, and for clinician education were raised as particularly important if tNGS is going to add value. The recommendation is that the data from the tNGS platform should first be carefully reviewed before it is communicated to clinicians, to provide a safety net that ensures the information is useful and that clinicians are not overwhelmed with data. In addition, a participant noted that experience has shown that clinicians would benefit from training to motivate for the effective use of the tNGS test results. This is of particular importance where there are no specialized MDR -TB doctors:

> *"When the LPA assay was first introduced, the doctors said: 'No, I'm not using this, this is not how I was trained, I was trained on phenotype, now I'm getting information on genotype. Now how do I handle that?'…Particularly when you don't have highly skilled MDR doctors everywhere…So, how are you going to equip the doctors to handle this report is something that needs to be considered. If it's a nursing Sister, you know, how does she deal with it. Those are all the challenges that lie ahead. This is not just 'plug and play." (Global expert 3).*

Discussion

The aim of this study was to gather qualitative evidence on implementation considerations of the new tNGS technology for diagnosing multi-drug resistance TB, to inform the new WHO guideline on rapid TB diagnostics. Against the background of a gap in qualitative evidence, we conducted a primary qualitative study to provide information on experiences and perceptions of stakeholders involved with testing out the tNGS, with respect to the acceptability and feasibility and other implementation considerations. Findings from the study can provide direct evidence to draw on, especially given the gap in published studies on this topic. We also sought the views of global TB experts, who provided broader insights on implementation consideration of new rapid molecular TB diagnostics.
A summary of the findings is provided in Table 1 below. A recent overview of the 'Implementability' of health care interventions concluded that scalability and sustainability of a healthcare intervention are dependent on its acceptability, fidelity, and feasibility, and that these factors should be considered at the intervention development stage and iteratively assessed throughout implementation (7).

Table 1: Summary of main and sub-findings

Main finding	Sub-findings
Acceptability There is an overwhelmingly positive sentiment for the acceptability and potential utility of tNGS technology, due to its comprehensiveness, convenience, and rapidness, as well as the timeliness of the	1. tNGS is a 'major advancement' in TB diagnostics 2. There is a good window of opportunity for utility of tNGS

test.	
Feasibility The overall sentiment was that the tNGS technology needs to be further developed before it can be considered fully ready for operational use.	1. Setting up the new technology 2. High technical complexity of the test 3. Specialized laboratory infrastructure and human resources 4. Specialist requirements for operating the test 5. Supply chain challenges 6. Data management and storage requirements 7. Continuous updating of the WHO mutations reference library 8. There are different feasibility concerns for the different tNGS platforms
Values, preferences, and equity The overall sentiment is that that MDR-TB diagnostic technology needs to balance accuracy, speed, affordability, equity, and cost effectiveness, and that tNGS technology would need to address these considerations before it can be implemented in these LMIC settings. These values, preferences and equity considerations were consistent across the different stakeholder groups who participated in the study.	1. Equitable access (centralized vs decentralized placement) 2. Affordability & cost-effectiveness 3. Synergies for optimizing diagnostic algorithms 4. Strengthening the broader laboratory and health system

Findings from the primary study highlighted the positive views of all stakeholders on the acceptability and value of tNGS technology, noting that this is major advance in rapid MDR-TB diagnostics. There was high praise for the capacity of tNGS technology to provide a diagnosis of drug resistance to a wide range of TB drugs, including for the newest TB drugs. The convenience of using a sputum sample (as compared to a culture sample), and the shortened turn-around time (3-5 days as compared to 4-6 weeks for phenotypic testing) are considered a major advance. The overall sense was that the diagnostic gains from tNGS technology has huge potential to make a positive impact on MDR-TB treatment outcomes, as it can allow for comprehensive diagnosis of drug resistance, in a timely fashion, and guide appropriate treatment of MDR-TB.
Nevertheless, implementers outlined several problems with testing the tNGS platforms, which would they felt would need to be addressed if the technology is to be fully acceptable and feasible for use in these LMIC settings. Challenges related to planning and setting up the platforms, including difficulties importing new and specialized equipment; highly specialist requirements for laboratory infrastructure and human resources that limits the use of the technology to centralized reference laboratories, and supply chain problems that jeopardized their need for continuous supply of specialist supplies they needed. There were concerns about limiting steps in the data management and storage components of the technology.

The high level of complexity of the tNGS technology was challenging even for specialist trained staff and left little room for error. The high complexity (and associated specialist infrastructure and human resources) has equity implications as it limits the placement of the technology to centralized labs. This may mean less even access for some regions within a country. Costs, cost effectiveness and equitable access to the technology for those most in need were key concerns. Finally, participants made recommendations on what they would ideally like to see in future iterations of tNGS and in other new rapid MDR-TB diagnostics. The wish is for affordable, point of care rapid MDR-TB diagnostics, that is

more automated and simplified ('plug and play'), to allow for decentralized laboratory use, closer to the point of diagnosing patients. There was a sense of caution about the need to rationalize and integrate new rapid diagnostics. There is also the concern that diagnostic gains from rapid tNGS technology, can only be optimized if there is investment in reengineering and strengthening laboratory systems and the broader health system.

As noted earlier, participants acknowledged that while start-up problems with the new technology would likely resolve over time, other feasibility challenges would likely remain unless the technology is further refined. For example, the lack of local, in-country supply chains would mean that challenges to specialized supplies and specialized technical support could remain, and affordability and cost-effectiveness considerations would likely require ongoing attention.
The challenges identified in this review are similar to the implementation considerations noted in the 2018 WHO guideline on tNGS use, especially those related to setting up of equipment and the workflow, infrastructure and human resource requirement and data analysis and storage (5).

Additional challenges highlighted in this primary study is the technical complexity of the tNGS procedures, especially the critical DNA library preparation steps, and the challenges to ensuring a continuous supply of specialist kits and reagents. Findings from this study correspond to those in a recent Cochrane review on user perspectives and experiences with low-complexity nucleic acid amplification tests (NAATs). NAATs are rapid tests for detection of tuberculosis and tuberculosis drug resistance used routinely in some settings. The Cochrane review found that "healthcare providers value having accurate tests that give them confidence in the diagnosis, rapid results, and keeping cost low, being able to use different specimens (such as sputum and stool) and receiving information about drug resistance as part of the test results. Laboratory personnel appreciated that laboratory work was made easier, and that staff was more satisfied thanks to rapid molecular diagnostic tests." (8). This echoes the findings from stakeholders in this primary study, with regards to appreciation for the value of rapid diagnostics. Other shared concerns between the review and this study were the implementation challenges in infrastructure, human resources, and supply chain systems and consideration of the ease of use (technical complexity) of testing.

Factors shaping implementation usually include the characteristics of the intervention or the technology itself, as well as organisational issues related to the human resources, infrastructure, health information, supply chain and financial systems (7). An important concern raised in this study that was echoed in the same Cochrane review, is that gains from rapid diagnostic testing may not be optimized in a poorly functioning health system. The Cochrane review identified challenges associated with diagnostic delays and underutilization of rapid TB diagnostics. They noted that "Delays were reported at many steps of the diagnostic pathway owing to poor sample quality; difficulties with transporting specimens; lack of sufficient resources; maintenance of low-complexity NAATs; increased workload; inefficient work and patient flows; over-reliance on low-complexity NAAT results in lieu of clinical judgement; and lack of data driven and inclusive implementation processes." (8).

Similarly, in this this primary study, stakeholders cautioned that the gains of rapid diagnostics can only be optimized if it supported with improvements in the functioning of the laboratory and broader health services and systems. Other points of overlap include concerns about affordability, sustainable funding, maintenance, and equitable access and use of about rapid diagnostics. Of interest, is that the Cochrane review also reports on obstacles to effective utilization of rapid TB diagnostics from the perspective of patients, which is an understudied area, noting that there is a need for programmes to support patient health literacy (8).

Implications for research and practice

As and when tNGS technology may be taken up in routine settings, it would be important to have comprehensive evaluations that can provide evidence of impact on broader TB treatment outcomes. Research using mixed methods would be useful to improve understanding of what works, when and why. Such research should include the perspectives health providers and patient groups for a comprehensive view of how to optimise diagnostic gains from tNGS. Policy and practice initiatives may

be required to strengthen laboratory and health system support functions, to allow for rational, integrated, and equitable use of tNGS technology.

Strength and limitations

The primary qualitative study responds to the gap in qualitative research on tNGS technology. To our knowledge, this is the first study on the experience and perceptions of implementers of tNGS for MDR-TB diagnostics. The findings from this study is timely as it can be used to guide further development of the tNGS technology. The strength of the study is that it was conducted close the implementation phase of FIND trial, which meant that experience and perceptions were still fresh in the minds of implementers. The participants represented a relatively comprehensive sample given it included that most in-country implementers that participated in the trial. The credibility of the findings is enhanced by the mix of participants that included the FIND Trial implementers, implementers elsewhere, and global experts, as this allowed for triangulation of opinions as well as for a broader range of perspectives to be explored. Participants were provided with the draft interview summaries to check for accuracy, and this quality control step also strengthens the credibility of the findings.
The downside of having multiple groups represented in a smaller study sample study like this, is that there is a trade-off between a broader scope of perspectives, but with less depth and detail in one's understanding of the different perspectives. Nevertheless, the perspectives shared here represent a fair range of insights raised elsewhere in the literature on factors shaping implementation of new interventions. There are also study limitations related to exploring such new technology. For instance, it is unclear the extent to which some of the experiences are related to newness of the technology mainly, and the extent to which those challenges might be resolved with time and with more routine uptake of the new technology. This applies specially to challenges related to set up and start-up, and to difficulties with ensuring a continuous supply chain of specialist products, and with quickly addressing technical hiccups. The scope of the investigation was also limited: as the technology was not yet tested for use in clinical decision-making, the study could not yet explore responses to the new technology amongst clinicians and patients.

Reflexiveness

The author is a social scientist with extensive experience in primary qualitative research as well as evidence synthesis methods. Implementing tNGS technology requires specialist laboratory skills and the author does not have specialist knowledge of laboratory science or tNGS technology. While specialist diagnostic expertise was not needed to conduct this research, a basic understanding of the tNGS technology was needed to explore the experience of participants. As the author is not a laboratory scientist, the first step was to acknowledge the gap in her knowledge and to take steps to educate herself on the tNGS technology. This was done through reading up on the topic and further consolidating her technical understanding through initial interviews with the FIND trial researchers. Nevertheless, the lack of specialised laboratory science expertise may have limited the richness of the more technical enquiry and findings. On the other hand, the author has broad health system research experience, including having researched implementation dynamics of new interventions like new rapid molecular TB diagnostics (GeneXpert). This health systems perspective helped to guide her exploration of experiences and perspectives of stakeholders in a more comprehensive way, which contributed to depth of the enquiry about the health systems implications of new technologies.

Overall implementation considerations

We extracted a set of implementation considerations based on the views and recommendations by participants in this study. There was a high level of acceptability of tNGS technology, and it is regarded as a major advance in rapid diagnoses for drug resistance in MDR-TB. However, a common sentiment was that technology needs to advance further to increase its acceptability, feasibility, and utility. The following are a range of issue for considerations by manufacturers of the tests and for health system, service, and policy decision-makers policy makers.

1. Planning, preparation, and implementation support will be needed to implement tNGS in LMIC settings.

a. Ensure the procurement of equipment is in place, including the required in country custom clearances.
b. Training programmes should allow sufficient opportunity for local staff to do multiple sequencing runs by themselves, under supervision of the trainers.
c. Include a checklist of challenges and errors to anticipate in the set-up and running of the sequencing platform, and instructions for trouble shooting.
d. Include a checklist of all specialist consumables and the quantities needed and who the suppliers are. Also include details of the general laboratory consumables needed.
e. Technical support from local vendors are required for technical support and maintenance of the tNGS platforms.

2. Specialized human resources are needed, and this will require planning and capacity development.
    a. Specialist molecular, medical scientists, with experience in molecular testing will need to be in place and they will need sufficient training on the tNGS platforms.
    b. Recruitment, renumeration and retention of specialist laboratory staff will need to be in place.
    c. Future iterations of tNGS technology should be aim for less complex testing technology (more automation, fewer complex processes), to allow for general laboratory personnel to implement the technology in decentralized laboratory settings.

3. Specialist laboratory infrastructure and supporting services need to be in place.
    a. Specialist laboratory infrastructure that is suitable for molecular testing is required, which limits the technology to centralized, reference laboratories. If the complexity and number of steps are reduced through semi-automation and standardization, and the laboratory infrastructure is reengineered, then tNGS technology could be suitable for use in decentralized laboratories.
    b. Additional infrastructure requirement should be specified upfront (e.g., number of rooms, air-conditioning requirements).
    c. Uninterrupted supply of electricity is required to effectively implement tNGS technology, and this is an important consideration in some LMIC settings.
    d. Internet connectivity of high capacity and speed will be needed to accommodate the large data files generated.
    e. Large computation capacity will be needed for data analysis and storage.

4. Uninterrupted supply chain for procurement of specialist consumables need to be in place.
    a. Preparation and planning is needed to ensure an uninterrupted supply of the specialist supplies.
    b. Procurement processes needs to accessible, simple, timely and affordable.
    c. This may require support from local vendors to ensure the appropriate custom clearances, and to remove supply chain blockages.
    d.

5. The tNGS platforms require improvements to the ease of use to improve its acceptability and feasibility, including lowering the technical complexity, providing more automation, and addressing data management, storage, and ownership issues.
    a. The tNGS platforms need to be field ready in terms of training and implementation protocols, and with the appropriate equipment in place.
    b. tNGS is a high complexity technology. Preparation of the sample prior to sequencing needs to be simplified. This will require reducing the number of steps, and more automation and standardization to increase efficiency and effectiveness of implementation.

c. There needs to be more of a seamless linkage between the sequencing technology and the data analysis technology of the tNGS platform.
d. Data analysis and data storage requirements need to be fully supported, including systems for backing up data.
e. Data ownership and data security considerations will need to be put in place.
f. Mechanisms are needed for interlinking data systems, to allow for two-way communication between the information system of the tNGS platform and the information system of the laboratory. This interlinking must be compatible with the internal security considerations of the laboratory information system.

6. Affordability and comprehensive cost-effectiveness appraisal is needed to inform rational use of tNGS.
    a. tNGS technology should be affordable for LMICs.
    b. Adoption and use of the tNGS platforms should include sustainable funding models.
    c. Costing should include capital costs of equipment (including replacement costs), as well as the costs of continuous supplies of specialist consumables, and of general laboratory consumables
    d. Costing should include cost of technical support and maintenance, preferably from local vendors.
    e. Cost effectiveness should include several contextual factors relevant to each country. These include the burden of disease, integration into the TB treatment algorithm, what alternative testing technologies are available, how to balance accuracy and speed of testing, and how to balance the cost of maintaining current services, with the cost of new expensive diagnostics.
    f. Cost effectiveness should include evidence of positive impact of tNGS technology on MDR TB treatment outcomes and ultimately on measures of improved TB control on population level.

7. Optimizing the gains of rapid diagnostics like tNGS depends on improvements in the function of the laboratory and the broader health systems.
    a. A well-functioning and appropriately funded laboratory system will strengthen the effective use of tNGS platform.
    b. Strengthening the linkages between the laboratory system and the rest of the health system will enhance the impact of the gains from rapid diagnostics. This includes an effective sample referral systems, and closer communication between laboratory staff and clinicians to ensure effective transfer of information for patient treatment.
    c. Patient diagnosis, referral, treatment and recall systems and drug availability will all need to be strengthened to optimise the gains from rapid diagnostics and to promote effective MDR-TB treatment.
    d. Education and training will be needed for clinicians on the use of test results from tNGS technology.
    e. Patient education, engagement and community empowerment is needed to spread awareness and encourage appropriate use of and benefit from rapid TB diagnostics.

Conclusion

While there is high acceptability of the value and potential of tNGS technology for MDR-TB management, a range of factors shaping acceptability, feasibility, and preferences of tNGS technology would need to be considered in future development of the technology. Guidelines on the use of tNGS technology, especially in LMICs, would need to take account of not only the effectiveness of the technology in diagnosing MDR-TB, but also the influences such as the ease of use of the technology, as organizational contexts and needs, and the values, preferences, and aspirations of stakeholders.

Contributions of authors

Dr Natalie Leon is the lead author of the protocol and the guarantor of the review.

Acknowledgements

The following individuals and groups are acknowledged for their contribution and support of this work. Participants in the interview study are thanked for their generosity in sharing their insights and their time. Experts are thanked for responding to the request for papers, and members of the WHO Expert advisory group are thanked for their comments on drafts of the work. Cecily Miller, Nazir Ismail, and Tamara Kredo as part of the WHO team, are thanked for technical support and advice.

Declarations of interest

Natalie Leon declare no financial or other (personal, political, academic) conflicts of interest.

References

1. Tiberi S, Utjesanovic N, Galvin J, et. al. Drug resistant TB – latest developments in epidemiology, diagnostics and management. International Journal of Infectious Diseases. 2022.
2. World Health Organization. Global Tuberculosis Report 2022. Geneva: World Health Organization; 2022.
3. Farhat M, Fresch L, Calderon R, Loerger T, Snyder M, Meehan C, et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nature Communications. 2019;10(2128).
4. World Health Organization. Global Tuberculosis Report 2016. Geneva: World Health Organization; 2016.
5. World Health Organization. The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide Geneva: World Health Organization; 2018. Report No.: Licence: CC BY-NCSA 3.0 IGO Contract No.: WHO/CDS/TB/2018.19.
6. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 3: diagnosis - rapid diagnostics for tuberculosis detection, 2021 update. Geneva: World Health Organization; 2021. Contract No.: Licence: CC BY-NC-SA 3.0 IGO.
7. Klaic M, Kapp S, Hudson P, Chapman W, Denehy L, Story D, et al. Implementability of healthcare interventions: an overview of reviews and development of a conceptual framework. Implement Sci. 2022;17(1):10.
8. Engel N, Ochodo EA, Karanja PW, Schmidt BM, Janssen R, Steingart KR, et al. Rapid molecular tests for tuberculosis and tuberculosis drug resistance: a qualitative evidence synthesis of recipient and provider views. Cochrane Database Syst Rev. 2022;4(4):CD014877.
9. Mounier-Jack S, Mayhew SH, Mays N. Integrated care: learning between high-income, and low- and middle-income country health systems. Health Policy Plan. 2017;32(suppl_4):iv6-iv12.
10. Lehmann U, Gilson L. Actor interfaces and practices of power in a community health worker programme: a South African study of unintended policy outcomes. Health Policy Plan. 2013;28(4):358-66.
11. Orgill M, Gilson L, Chitha W, Michel J, Erasmus E, Marchal B, et al. A qualitative study of the dissemination and diffusion of innovations: bottom up experiences of senior managers in three health districts in South Africa. Int J Equity Health. 2019;18(1):53.
12. Walker L, Gilson L. 'We are bitter but we are satisfied': nurses as street-level bureaucrats in South Africa. Soc Sci Med. 2004;59(6):1251-61.
13. Erasmus E. The use of street-level bureaucracy theory in health policy analysis in low- and middle-income countries: a meta-ethnographic synthesis. Health Policy Plan. 2014;29 Suppl 3:iii70-8.
14. Erasmus E, Gilson L. How to start thinking about investigating power in the organizational settings of policy implementation. Health policy and planning. 2008;23: 361-68.

15. Youssef A, Chaudhary ZK, Wiljer D, Mylopoulos M, Sockalingam S. Mapping Evidence of Patients' Experiences in Integrated Care: A Scoping Review. Gen Hosp Psychiatry. 2019;61:1-9.

Appendix 1: Interview guide for primary qualitative study

QUALITATIVE RESEARCH ON EXPERIECES AND PERCEPTIONS WHO TNGS RAPID DIAGNOSTICS INTERVIEW GUIDE (Draft, V3). 9 Sept 2022
LABORATORY STAFF AND MANAGEMENT IN TRIAL STUDY SITES

Introductions

- Aims of the interview
- Consent procedures
- Questions

Interview questions:

a) Your role and responsibilities

1. Please could you tell me about what your role and responsibilities at this institution.
    a. How long have you been doing this work? And in this laboratory?

b) Background and experience with molecular and sequencing technologies

2. What types of molecular and sequencing technologies are you currently using in this laboratory?
    a. For what purpose are these molecular and or sequencing technologies used- research, surveillance, patient management?
    b. Do you have prior experience with molecular and or sequencing techniques? What does the experience comprise?

c) Background and experience with the tNGS TB rapid diagnostic technology

3. When and how did you become involved with the trial study to test the tNGS technology?
    a. What is the name of the tNGS technologies you are using in this trial study setting?
    b. What is your role in the trial study and in implementing of the tNGS technology?
    c. Do you have prior experience with NGS or tNGS? What does the experience comprise?

4. From your perspective as a laboratory professional, in the laboratory setting (in the context of this trial study):
    a. What would you say are the objectives of the tNGS technology you are using?
    b. And the benefits and drawbacks of the tNGS technology you are using?
    c. Thinking ahead to its clinical application, what do you think the potential benefits and drawbacks are for clinical management of TB patients?

d) Understanding the tNGS rapid diagnostic process that is being used in the trial study setting

I would like to get a clear view of the tNGS workflows you use in your laboratory setting. With reference to your role and the role of the rest of the team:

5. Could you please take me through the steps from receiving samples for testing to receiving the reporting of test results and describe the implementation steps and considerations.
6. How does this process flow differ from other NGS technologies you may be using in this laboratory?
7. What were the training and supervision requirements to implement this tNGS technology? And how did that go?
8. What other preparatory work and resources were required? And how did that go?

e) Understanding your experience and perceptions of the tNGS technology

I would now like to explore in some detail, what your general experience is of the tNGS technology you are testing. Looking at the full continuum of the testing process: Phase I: sample prep and library prep 2) sequencing 3) data analysis and reporting, what has the experience been like for you?

a) What has worked well? And why?

b) What has not worked well? And why?

9. For each of the tNGS solutions? What was the experience like for you?
    a. What has worked well? And why?
    b. What has not worked well? And why?

10. How would you summarise what the requirements were for implementing each of the tNGS solutions?
    a. What the challenges were
    b. And how have you team responded to those challenges.

11. How does your experience of tNGS with each of the three solutions compare with other DR-TB diagnostics you may have been involved with/ and or know of, such as liquid culture, or LPA?
12. How does your experience compare with other rapid DST- TB diagnostics in this laboratory/health service setting?
13. How does your experience compare with phenotypic DST TB diagnostics in this laboratory/health service setting?
14. How does your experience compare with diagnostics for DST for treatments other than TB (e.g., HIV antiretroviral treatment) in this laboratory/health service setting?

f) Overview of l experience and perceptions (if not already covered sufficiently above)

15. What is your sense of the potential value of tNGS rapid diagnostics for TB care (in general and in your country contexts)?
16. What is your sense of the main limitations of tNGS rapid diagnostics for TB care (in general and in your country context?)
17. What if any concerns do you have about the appropriateness and acceptability of the tNGS rapid diagnostic technology in your country context?
18. What if any concerns do you have about the feasibility of the tNGS rapid diagnostic technology in your country context?
19. What if any concerns do you have about the equity of the tNGS rapid diagnostic technology in your country context? For example, with reference to the extent to which the benefit of the technology reaches those who need it most?
20. What would need to change in your country context to enable the tNGS technology to have an optimal positive impact on clinical care for TB patients. What is your sense of how likely (feasible) is it that these enabling factors will be put in place in your setting?

g) Questions from the interviewee and concluding comments & thanking the participant

For more information, contact Natalie Leon. natalieheleneleon@gmail.com

Appendix 2: Information and consent form

EXPERIENCES AND PERCEPTIONS OF TARGETTED NEXT GENERATION SEQUENCING (tNGS) RAPID DIAGNOSTICS FOR DETECTING TB DRUG RESISTANCE: AN INTERVIEW STUDY TO INFORM THE WHO tNGS GUIDELINE

Information and consent form (V3, 2022)

Study background and rationale

Targeted New Generation Sequencing (tNGS) is a novel, rapid molecular diagnostic test that provide an end-to-end solution for detecting resistance to multiple TB drugs at the same time, but little is known about its implementation considerations. To inform the World Health Organization (WHO) guideline on tNGS use for testing multi drug resistance in TB, this study will explore stakeholder views on acceptability, feasibility, equity, values, and preferences of using tNGS in the three FIND trial testing sites, and in other settings using tNGS.

Dr Natalie Leon will be conducting the interviews. She is an independent researcher contracted by WHO to gather qualitative evidence on implementation considerations of tNGS. She is contactable at natalieheleneleon@gmail.com and on WhatsApp +1 8042451481.

Why you were approached for an interview

- You were approached to participate due to your research, practice, clinical or policy experience with tNGS for drug resistance testing and/or similar rapid diagnostic experience. Sharing your experience and perceptions of the tNGS technology will provide useful information on implementation considerations to inform the WHO guideline.

What is required from you?

- The interview will be maximum 1 hour and will explore your experience and perceptions of implementing the tNGS rapid testing technology, especially in Low-and middle-income (LMIC) settings.

Ethics, Confidentiality and Consent

- A protocol for the scope and methods of the qualitative study was submitted and agreed to by the WHO team who commissioned the study. (Contact WHO representative, Dr Cecily Miller cmiller@who.int for more information).
- Ethical oversight of this work has been reviewed by the methodologist and WHO Secretariat and is sufficient for the information being collected.
- You will be asked to give written consent for the interview, by signing this consent form.
- The information you share is kept confidential in that it is synthesized and anonymised, so your name will not be linked to the data. The information will form part of a report to WHO in Nov 2022 to inform the WHO guideline, on key emerging implementation issues that emerged across views of multiple stakeholders.

Benefits and harms

- We do not anticipate any benefits or harms to you as a participant in this interview. Your participation will contribute information to inform the WHO guidelines.
- You are free to decline questions or to stop the interview.
- For information or concerns, please contact the WHO representative, Dr Cecily Miller cmiller@who.int

I agree for this interview to be recorded*: Please pick one:* YES or NO

I (*name and surname*) ……., give my consent to participate in this interview.

Signature ……. Date: / / 2022

Paste signature/write initials, then save as PDF and email or print, sign and email scanned printed document to natalieheleneleon@gmail.com

Web Annex D.21. Stakeholder experiences and perspectives of using Targeted next generation sequencing (tNGS) for diagnosing multi-drug resistant Tuberculosis (MDR-TB): a qualitative evidence synthesis

(Draft, 15 December 2022, Final draft 17 April 2023)

Author:
Dr Natalie Leon
Independent public health research consultant
nataliehelenleon@gmail.com

A project for:
Unit: TB Prevention, Diagnosis, Treatment, Care & Innovation (PCI)

Abstract

BACKGROUND

Timely drug resistance detection is essential to global tuberculosis management. Novel molecular testing tools are being developed to rapidly detect multi-drug resistant Tuberculosis (MDR-TB) which is needed for more effective and timely treatment of MDR-TB. Targeted next generation sequencing (tNGS), is new molecular-based rapid diagnostic technology (used on a sputum sample), that can provide comprehensive diagnosis of MDR-TB in a short turnaround time of a few days, compared to the standard culture-based phenotypic MDR-TB diagnostics that take multiple weeks. There are, however, challenges to effective implementation and wide scale uptake of rapid diagnostics, including technical logistical, organisational, economic, and contextual factors. There is a need to understand how these factors may be shaping implementation. Exploring the experiences and perceptions of health workers implementing rapid TB diagnostics can provide useful information to better understand the implementation factors shaping the use and impact of rapid diagnostics.

OBJECTIVE

This study aimed to provide evidence to inform the formation of new WHO guideline for use of tNGS technology for MDR-TB diagnosis. The objective is to identify and synthesize qualitative evidence on tNGS for MDR-TB diagnostics, to examine the implementation considerations related to acceptability, feasibility, values and preferences and equity, from the perspectives of health workers as implementers.

METHODS

This was a rapid review of qualitative evidence, guided by the rapid review recommendations for streamlining methods used for systematic reviews. The review steps included electronic and open searching and approaching experts to identify records. and reviewing records. We searched Medline with no year or language limits. We designed a search strategy and ran on Aug 19, 2022, and rerun on Oct 10, 2022, to include WGS related studies for MDR-TB. Records identified through the search were downloaded into an EndNote file and abstracts were screened for relevance by a single author. Eligible abstracts were identified, the full text papers were retrieved and screened for relevance by a single author. During the full-text review stage, also be a single reviewer, the author consulted with the WHO team on several full-text papers to get a second opinion. Where eligible papers were to be included for data analysis, data would be extracted for the main areas of interest and thematically synthesized. Studies eligible for inclusion would be screened for quality to consider contribution to overall certainty of findings, as part of a CERQual assessment of findings.

RESULTS

The review did not identify any eligible studies for analysis and synthesis. Based on the systematic electronic search, we identified 3 records. Based on the open, hand and expert search, we found 27 records. On full-text review of the 30 records, none were found to be eligible for inclusion. The main reason for exclusion was on study method (quantitative or discussion papers), or the study focus was out of scope (for example, not based on implementation experience, but on hypothetical views). The PRISMA diagram in Figure 1 shows the flow diagram for the search and screening process.

DISCUSSION

Acceptability, fidelity, and feasibility considerations are considered key for implementation and sustainability of interventions. This study sought to identify and synthesize qualitative evidence on acceptability, values, preferences, equity, and feasibility of tNGS rapid diagnostics for MDR-TB. The review did not yield any direct evidence from primary qualitative perception studies on tNGS technology as no studies could be found. Given this is an empty review, there are clearly research gaps waiting to be filled. While there are several studies testing the accuracy of tNGS technology for MDR-TB diagnostics in controlled settings, the technology is not yet widely operational, which may explain why there are no primary perceptions studies about implementation of tNGS.

In the absence of direct evidence on tNGS for MDR-TB diagnostics, it is recommended that WHO guideline developers and policy decision-makers draw on insights from the primary research study on the experience and perceptions of participants who implemented the tNGS technology as part of the FIND trial (Leon 2023 report for WHO). A recent (2022) Cochrane review of perceptions of health providers and patient recipients of other rapid TB diagnostics can provide useful indirect evidence to

inform the new WHO guideline on tNGS technology. Of interest, is that the Cochrane review also reports on obstacles to effective utilization of rapid TB diagnostics from the perspective of patients as recipients of rapid diagnostics, and they recommend programmes to support patient health literacy. Patient recipients' experiences and perspectives of rapid TB diagnostics is an understudied area that deserves more attention.

CONCLUSION
A rapid review of qualitative evidence on the experiences and perceptions of tNGS technology for MDR-TB diagnostics was conducted to inform the development of a new WHO guidelines on tNGS technology. The review did not yield any eligible records for analysis of issues related to acceptability, feasibility, values, preferences and equity of the tNGS technology, most likely due to the newness of the tNGS technology in routine settings. Going forward, more qualitative studies are needed alongside trials and operational implementation of tNGS for MDR-TB.

Background

Tuberculosis (TB) remains a major global public health threat. One of the most challenging forms of the disease is multidrug resistant TB (MDR-TB), due to higher morbidity and mortality, complexity of treatment and higher cost (1). WHO estimated that in 2016, close to half a million people were diagnosed with MDR-TB (2). It is estimated that only two out of every three patients with MDR-TB are diagnosed, of those diagnosed, only three out of four are treated, and only half of those treated are cured. This results in the majority (75%) of patients with MDR-TB not being cured; thus, persisting with their illness, spreading the disease and/or dying from their illness. Reasons for this loss of engagement in care include health service factors such as diagnostic delay, treatment delay, inaccurate treatment and patient delays in seeking health care (3).
Timely drug resistance detection is essential to global tuberculosis management (4). The End TB Strategy of the World Health Organization calls for the early diagnosis of TB and universal drug-susceptibility testing (4). Traditional drug susceptibility testing uses culture-based tests which can take several weeks to yield a result and may not be widely available for testing for newer and repurposed drugs (1). Novel molecular testing tools are being developed to rapidly detect TB and resistance to anti-TB drugs, and to diagnose resistance to multiple different TB drugs simultaneously. However, there are obstacles to its wide-scale uptake and implementation, especially in low-and middle-income (LMIC) countries where these diagnostics may be most needed. Obstacles relate to a range of technical, logistical, organizational, economic and contextual factors. There is a need to understand how these factors may be shaping implementation. Exploring the experiences and perceptions of health workers implementing rapid TB diagnostics can provide useful information to better understand the implementation factors shaping the use and impact of rapid diagnostics.
Genotypic testing, such as Whole genome sequencing (WGS) is molecular-based gene sequencing technology for use in screening and diagnosis of genetic disorders, cancers, as well as in diagnosis of drug resistance in infectious diseases such as HIV, malaria, and TB. Targeted next generation sequencing is a further advancement on WGS, where the focus is on a targeted section of the genome (rather than the whole genome). tNGS for MDR-TB diagnostics is molecular technology that is focused on identifying resistance to TB drugs by targeting investigation of the section of the genome that is known to be associated with mutations associated with resistance to TB drugs. Use of tNGS in TB diagnostics therefore has the capability to provide more comprehensive diagnosis of resistance to a wide spectrum of TB drugs as compared to current rapid diagnostics that identify resistances for a smaller number of TB drugs. tNGS technology can identify comprehensive multidrug resistance susceptibility in a shorter time (estimated at between 3 and 5 days), and on a sputum sample, as compared to the standard culture-based phenotypic diagnostic testing that takes several weeks (3 to 6 weeks). tNGS technology also provides testing at higher volumes which makes it suitable for use in high TB burden settings. tNGS platforms are aimed at providing an integrated end-to-end solution for MDR-TB drug resistance testing, including automated data analysis and reporting. Diagnostic reports can provide drug resistance information for each sample that was sequenced in a batch, that details which drugs the patient is resistant to (or not). Some platforms are also able to provide information on the lineage of mutations, which can be used to identifying and track strains of drug resistance. tNGS technology therefore offers the potential to provide a comprehensive diagnosis of multi-drug

resistance TB. tNGS diagnostic test results would enable timely clinical decisions on appropriate treatment for MDR-TB, which can improve TB treatment outcomes and stop the transmission of resistant strains of TB (5).

Different tNGS platforms have different laboratory workflows, but the general tNGS workflow includes: preparation of the TB sputum sample, DNA extraction from the sputum sample and DNA purification, library preparation, sequencing and data analysis, as shown in the workflow Diagram 1 below (5). DNA extraction procedures require specialized laboratory safety levels (biosafety level 3) and special safety equipment and work procedures for contamination containment. Following DNA extraction, DNA library is prepared that involved DNA fragmentation and enrichment, a complex process where it is critical for technicians to exactly follow the instructions of the manufacturers, in terms of use of reagents, controls of temperature and time, and with quality and quantity check required before and after the library preparation. The prepared sample is then run on a sequencing machine and there are different commercially available models. Computational resources are needed for data analysis, report generation and data storage (5).

Diagram 1: tNGS workflow diagram (5)

Targeted tNGS therefore hold the potential, especially in high-burden TB settings, for effective treatment and control of the spread of MDR-TB, but there are obstacles to the uptake of novel, rapid technologies. Obstacles to similar rapid diagnostics have been documented and include concerns about costs, integration into existing laboratory workflows, technical training and skill requirements for utilization of the technology, and the need for expert guidance regarding the management and clinical interpretation of sequencing data (6).

A recent overview of the 'Implementability' of health care interventions concluded that acceptability, fidelity, and feasibility of interventions may influence uptake and scalability and suggested that these factors be considered at the early stages of intervention development and during implementation and evaluation of interventions (7). Stakeholder views of new rapid TB diagnostics are important for informing implementation plans and for understanding the role of stakeholders in shaping implementation and outcomes (8). Health workers play an important role in shaping implementation and outcomes and may be influenced by their experience and perceptions of individual, organizational and system level factors (9-13). Some argue that health workers ultimately determine how interventions are implemented, based on their understanding of their task, and shaped by their discretionary power in delivering the task (14), hence the importance of understanding implementation considerations from their perspective. For instance, health workers as frontline implementers in low resource settings may be struggling with issues such as chronic staff shortages, multiple demands, and poor performance management, while in high income settings, the high levels of specialization and financial disincentives may shape engagement (9).

It would therefore be useful to better understand how new rapid diagnostic technologies like tNGS are experienced and perceived by those who are implementing the technology. It would be valuable to understand the experiences and perceptions of laboratory staff who are frontline implementers testing out rapid TB diagnostics, for insights on implementation considerations - such as the feasibility, acceptability, equity, and value preferences of new technology (8, 15). The view of TB patients and patient advocates on rapid TB diagnostics can also provide valuable insights, and this would be important to investigate once the new technology is taken up as part of standard care (8).

In preparation for the new WHO guideline development meeting on tNGS technology for diagnosis of MDR-TB (5), there is a need to identify and summarize the current evidence on the implementation considerations for use of tNGS. Given the newness of the use of tNGS for TB care, it was unclear if there would be qualitative evidence available for this evidence synthesis. Hence this qualitative evidence synthesis study will be likely be supplemented with a primary qualitative study on the experiences and perceptions on health workers who are implementing tNGS. A supplementary primary qualitative study is feasible given that there is the FIND trial underway testing tNGS in three countries: India, Georgia, and South Africa, with a goal of providing high-quality evidence of diagnostic accuracy of the tNGS technology. A report on the primary qualitative study will be presented separately (Leon 2023, April, submitted to WHO team).

Objective

This study seeks to provide evidence to inform the new WHO guideline for use of tNGS technology for MDR-TB diagnostics. The objective is to identify and synthesize qualitative evidence on implementation considerations of tNGS technology for MDR-TB, related to acceptability, feasibility, values and preferences and equity, from the perspectives of health workers as implementers.

Methods

Overall study design

This was a rapid review of qualitative evidence, guided by the rapid review recommendations for streamlining methods used for systematic reviews (16). Rapid review techniques balance the need for timely results with a commitment to maintaining the robustness, meaningfulness, transparency, and trustworthiness of the findings. Rapid review recommendations include limiting the number of electronic data bases for searching, using dual or single reviewers for the stages of abstract screening, full text reviewing, data extraction and synthesis, and for assessing the quality of studies (16). We developed a protocol guide that guided the stepwise process of searching and screening records for relevance, reviewing eligible full-text records and extracting data relevant to the questions of interest, and then synthesizing the data.

Inclusion and exclusion criteria

Types of studies

We included primary studies that used qualitative study designs such as cross-sectional observational studies, ethnography, case studies, and qualitative process evaluations. Studies were eligible if they used both qualitative methods for data collection (e.g., focus group discussions, individual interviews, observation, diaries, document analysis, open-ended survey questions) and qualitative methods for data analysis (e.g., thematic analysis, framework analysis, grounded theory). Mixed methods studies were eligible if they collected and analysed qualitative data.

Topic of interest

The focus of the review is on tNGS as a rapid diagnostic for drug resistance for MDR TB. We included Whole genome sequencing (WGS) if it focused on MDR-TB diagnostics for TB care. WGS is a molecular based test that can also be used for MDR-TB diagnostics and that has been used for longer, so it could potentially provide indirect evidence. We defined TB care services as health care aimed at detection, treatment, and cure of TB in children and adults. Studies were eligible if the primary focus was the experiences and attitudes of stakeholders on tNGS for MDR-TB diagnosis and treatment.

Type of intervention and settings

The tNGS technology is mainly used in reference laboratory settings. Reference laboratories are usually centralized laboratories within the public health system of a country. There are also public - private and private reference laboratories in some settings, which are sometimes found at tertiary and academic hospitals and research laboratories.

Types of participants

The focus was on the experiences and attitudes of stakeholders involved with the planning and implementation of tNGS and or WGS for MDR -TB in any country. Eligible participants included policy makers and programme managers, laboratory staff and management, frontline clinical managers, and staff. Given the newness of tNGS technology, we anticipated that perception studies would likely focus on the experience of health care staff and management who implemented the tNGS technology. Nevertheless, we would include studies on other stakeholders if these were available, such as, lay health workers, patients and their families and patient advocates.

Exclusion criteria

We excluded studies that did not have stakeholder experience of tNGS or WGS for MDR-TB care as the primary focus.

Search methods for identification of studies

Electronic searches

We used a combination of a systematic search strategy and an open searching approach. The search strategy used a set of search terms associated with NGS technology as an intervention, with MDR-TB and qualitative research (and associated names). We searched Medline with no year or language limits. We designed a search strategy and ran on Aug 19, 2022, and rerun on Oct 10, to include WGS related studies for MDR-TB. (See Appendix 1 for the search strategy terms). We complemented this search with open searching on Google and Google Scholar using combinations of key terms, as well as hand searching (following references within papers). We wrote to experts to ask if they had knowledge of potentially eligible studies.

Data management

Records identified through the search were downloaded into an EndNote file and abstracts were screened for relevance by a single author. Eligible abstracts were identified, the full text papers were retrieved and screened for relevance by a single author. During the full-text review stage, also by a single reviewer, the author consulted with the WHO team on several full-text papers to get a second opinion. Where eligible papers were to be included for data analysis, data would be extracted for the main areas of interest and thematically synthesized. Studies eligible for inclusion would be screened for quality to consider contribution to overall certainty of findings, as part of a CERQual assessment of findings.

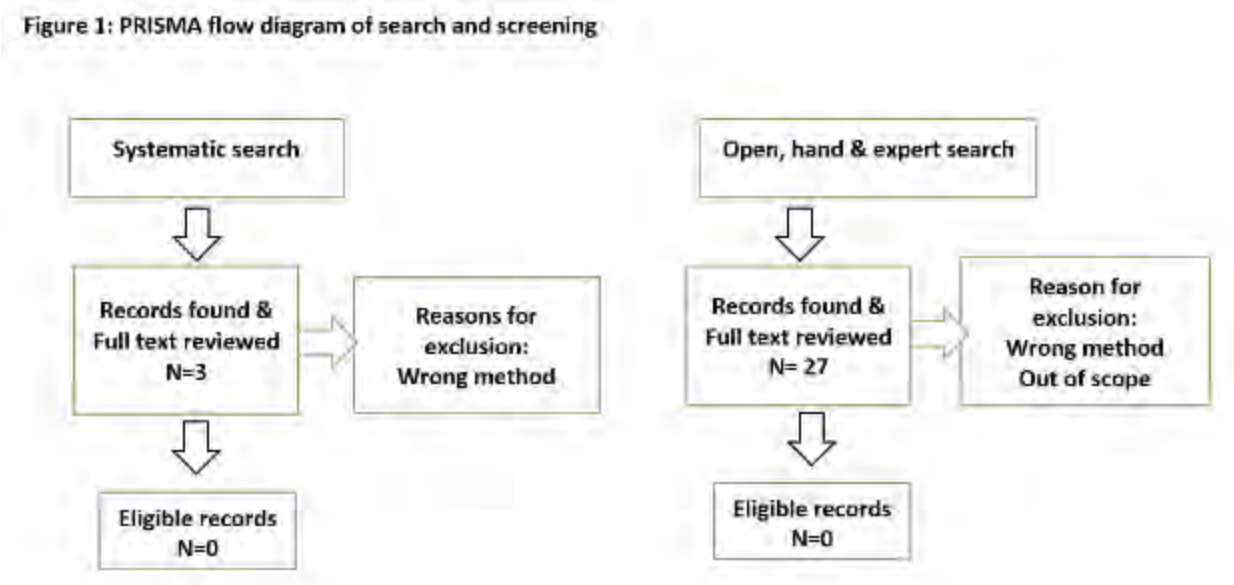
Ethical approval

Institutional ethical approval is not required for evidence synthesis research.

Findings

Based on the systematic electronic search, we identified 3 records. Based on the open, hand and expert search, we found 27 records. On full-text review of the 30 records, none were found to be eligible for inclusion. The main reason for exclusion was on study method (quantitative or discussion papers), or the study focus was out of scope (for example, not based on implementation experience, but on hypothetical views). The PRISMA diagram in Figure 1 shows the flow diagram for the search and screening process.

Figure 1: PRISMA flow diagram of search and screening



This was a rapid review of qualitative evidence on the perceptions and experiences of stakeholders on use of tNGS technology for diagnosing drug resistance in MDR-TB. The review did not identify any eligible studies for review and synthesis. The main reason for exclusion was on study method (quantitative or discussion papers), or the study focus was out of scope (for example, not based on implementation experience, but on hypothetical views).

Discussion and recommendations

The aim of this study was to identify and synthesize qualitative evidence on implementation considerations of the new tNGS technology for diagnosing multi-drug resistance TB, to inform the new WHO guideline on rapid TB diagnostics. This review did not yield any direct evidence from primary qualitative perception studies on tNGS technology as no studies could be found.

Research and practice implications

Given this is an empty review, there are clearly research gaps waiting to be filled. While there are several studies testing the accuracy of tNGS technology for MDR-TB diagnostics in controlled settings, the technology is not yet widely operational, which may explain why there are no perception studies about implementation in operational settings. Going forward, more effectiveness studies of the tNGS platforms are likely to be launched in multiple settings. This may provide useful opportunities to learn more about the experience and perceptions of implementers in different settings. It is recommended that researchers consider including parallel process evaluation studies when testing the accuracy of tNGS technology in operation settings, as this can shed light on the implementation challenges, from the perspective of the frontline implementers and managers. As and when tNGS technology may be taken up in routine settings, it would be important to have comprehensive evaluations that can provide evidence of impact on broader TB treatment outcomes. Research using mixed methods would be useful to improve understanding of what works, when and why. Such research should include the perspectives health providers and patient groups for a comprehensive view of how to optimize diagnostic gains from tNGS.

In the absence of direct evidence on tNGS for MDR-TB diagnostics, it is recommended that WHO guideline developers and policy decision-makers draw on insights from the primary research study on the experience and perceptions of participants who implemented the tNGS technology as part of the FIND trial (Leon 2023 report submitted April to WHO team). A potentially useful source of indirect evidence for the WHO guideline development process, is the 2022 Cochrane qualitative review on rapid molecular tests for TB drug resistance that have been in place for longer (such as low-complexity nucleic acid amplification tests (NAATs)) (8).

The recent 2022 Cochrane review is titled “Rapid molecular tests for tuberculosis and tuberculosis drug resistance: a qualitative evidence synthesis of recipient and provider views”, and it reports on stakeholder experiences and perceptions of various rapid TB diagnostics. The review identified several findings that may be of relevance to other rapid molecular tests for MDR-TB, such as tNGS. The review found that “healthcare providers value having accurate tests that give them confidence in the diagnosis, rapid results, and keeping cost low, being able to use different specimens (such as sputum and stool) and receiving information about drug resistance as part of the test results.” (8). The fit of the technology with current process flows were mentioned to be relevant: “Laboratory personnel appreciated that laboratory work was made easier, and that staff was more satisfied thanks to rapid molecular diagnostic tests.” (8) Other considerations in this review included setting up of equipment, the workflow, infrastructure and human resource requirements and data analysis and storage, issues that were also raised in in the WHO 2018 guideline that provided an overview of tNGS technology (5). The Cochrane review also identified challenges associated with diagnostic delays and underutilization of rapid TB diagnostics in various health system settings. They noted that “Delays were reported at many steps of the diagnostic pathway owing to poor sample quality; difficulties with transporting specimens; lack of sufficient resources; maintenance of low-complexity NAATs; increased workload; inefficient work and patient flows; over-reliance on low-complexity NAAT results in lieu of clinical judgement; and lack of data driven and inclusive implementation processes.” (8) Of interest, is that the Cochrane review also reports on obstacles to effective utilization of rapid TB diagnostics from the perspective of patients as recipients of rapid diagnostics, and they recommend programmes to support patient health literacy. Patient recipients’ experiences and perspectives of rapid TB diagnostics is an understudied area that deserves more attention.

Strength and limitations

The QES found no eligible studies most likely due the newness of the tNGS technology. We anticipated finding very little information, and a pilot search prior to embarking on the review may have confirmed the lack of information. The value of having conducted the review and coming up empty, it that it confirms there is gap in qualitative studies on implementation of TNGS for MDR-TB. This knowledge is useful to encourage researchers to include qualitative and or mixed method process evaluations alongside future operational research on tNGS technology.

The author is a social scientist with extensive experience in evidence synthesis methods, as well as primary qualitative research. The review used multiple search methods to identify potential studies, including open and hand searching, and contacting experts in the field, as we anticipated that published literature may be limited. Rapid review approaches have potential methodological limitations, given that some processes are less robust than in a full systematic review. For example, the use of a single reviewer at the screening and full-text review stage in this review holds potential for biase. To limit the potential for biase, the reviewer consulted with members of the WHO team during full-text screening when there was uncertainty.

Conclusion

A rapid review of qualitative evidence on the experiences and perceptions of tNGS technology for MDR-TB diagnostics was conducted to inform the development of a new WHO guidelines on tNGS technology. The review did not yield any eligible records for analysis of issues related to acceptability, feasibility, values, preferences and equity of the tNGS technology, most likely due to the newness of the tNGS technology in routine settings. More qualitative studies are needed alongside trials and operational implementation of tNGS for MDR-TB. A recent (2022) WHO commissioned primary qualitative study of experience and perceptions of implementers in the FIND trial can provide initial insights to inform the WHO guideline. A recent (2022) Cochrane review of perceptions of health providers and patient recipients of other rapid TB diagnostics can provide useful indirect evidence to inform the new WHO guideline on tNGS technology.

Contributions of authors

Dr Natalie Leon is the lead author of the protocol and the guarantor of the review.

Acknowledgements

The following individuals are acknowledged for their contribution and support of this work. Participants in the interview study are thanked for their generosity in sharing their insights and their time. Experts are thanked for responding to the request for paper, and members of the WHO Expert advisory group are thanked for their comments on the work. Cecily Miller, Nazir Ismail and Tamara Kredo as part of the WHO team, are thanked for technical support and advice. Christopher Colvin is thanked for assisting with designing the systematic search strategy and running the search.

Declarations of interest

Natalie Leon declare no financial or other (personal, political, academic) conflicts of interest.

References

1. Tiberi S, Utjesanovic N, Galvin J, et. al. Drug resistant TB – latest developments in epidemiology, diagnostics and management. International Journal of Infectious Diseases. 2022.
2. World Health Organization. Global Tuberculosis Report 2022. Geneva: World Health Organization; 2022.
3. Farhat M, Fresch L, Calderon R, Loerger T, Snyder M, Meehan C, et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nature Communications. 2019;10(2128).
4. World Health Organization. Global Tuberculosis Report 2016. Geneva: World Health Organization; 2016.
5. World Health Organization. The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide Geneva: World Health Organization; 2018. Report No.: Licence: CC BY-NCSA 3.0 IGO Contract No.: WHO/CDS/TB/2018.19.
6. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 3: diagnosis - rapid diagnostics for tuberculosis detection, 2021 update. Geneva: World Health Organization; 2021. Contract No.: Licence: CC BY-NC-SA 3.0 IGO.
7. Klaic M, Kapp S, Hudson P, Chapman W, Denehy L, Story D, et al. Implementability of healthcare interventions: an overview of reviews and development of a conceptual framework. Implement Sci. 2022;17(1):10.
8. Engel N, Ochodo EA, Karanja PW, Schmidt BM, Janssen R, Steingart KR, et al. Rapid molecular tests for tuberculosis and tuberculosis drug resistance: a qualitative evidence synthesis of recipient and provider views. Cochrane Database Syst Rev. 2022;4(4):CD014877.
9. Mounier-Jack S, Mayhew SH, Mays N. Integrated care: learning between high-income, and low- and middle-income country health systems. Health Policy Plan. 2017;32(suppl_4):iv6-iv12.
10. ehmann U, Gilson L. Actor interfaces and practices of power in a community health worker programme: a South African study of unintended policy outcomes. Health Policy Plan. 2013;28(4):358-66.
11. Orgill M, Gilson L, Chitha W, Michel J, Erasmus E, Marchal B, et al. A qualitative study of the dissemination and diffusion of innovations: bottom up experiences of senior managers in three health districts in South Africa. Int J Equity Health. 2019;18(1):53.
12. Walker L, Gilson L. 'We are bitter but we are satisfied': nurses as street-level bureaucrats in South Africa. Soc Sci Med. 2004;59(6):1251-61.
13. Erasmus E. The use of street-level bureaucracy theory in health policy analysis in low- and middle-income countries: a meta-ethnographic synthesis. Health Policy Plan. 2014;29 Suppl 3:iii70-8.
14. Erasmus E, Gilson L. How to start thinking about investigating power in the organizational settings of policy implementation. Health policy and planning. 2008;23: 361-68.
15. Youssef A, Chaudhary ZK, Wiljer D, Mylopoulos M, Sockalingam S. Mapping Evidence of Patients' Experiences in Integrated Care: A Scoping Review. Gen Hosp Psychiatry. 2019;61:1-9.
16. Tricco AC, Langlois EV, Straus SE. Rapid reviews to strengthen health policy and systems: a practical guide. Geneva: Alliance for Health Policy and Systems Research & World Health Organization; 2017. Contract No.: License: CC BY-NC-SA 3.0 IGO.

Appendix 1: Search strategy for QES

Database: Ovid MEDLINE(R) ALL <1946 to August 19, 2022>
Search Strategy:
---
1 "targeted next generation sequencing".mp. (2890)
2 "targeted NGS".mp. (782)
3 1 or 2 (3284)
4 Interviews as Topic/ or interview*.mp. or Interview/ (458448)
5 survey*.mp. or Health Surveys/ or Health Care Surveys/ or "Surveys and Questionnaires"/ (1192334)
6 Qualitative Research/ (76141)
7 Focus group discussion*.mp. or Focus Groups/ (42217)
8 "mixed methods".ti. or "mixed methods".ab. or "mixed-methods".ti. or "mixed-methods".ab. (27029)
9 4 or 5 or 6 or 7 or 8 (1563565)
10 3 and 9 (23)
11 Extensively Drug-Resistant Tuberculosis/ or Tuberculosis/ or tuberculosis.mp. [mp=title, book title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (270447)
12 Tuberculosis, Multidrug-Resistant/ or Tuberculosis, Pulmonary/ or Mycobacterium tuberculosis/ (125151)
13 (Tuberculosis or MDR-TB or XDR-TB or tuberculous).ti. (174344)
14 (Tuberculosis or MDR-TB or XDR-TB or tuberculous).ab. (132667)
15 11 or 12 or 13 or 14 (275832)
16 10 and 15 (3)

****

For further information, please contact:
Global Tuberculosis Programme
World Health Organization
20 Avenue Appia CH-1211 Geneva 27 Switzerland
Web site: https://www.who.int/teams/global-tuberculosis-programme/overview

9789240090477