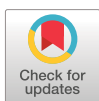# Clinical utility of WHO-recommended screening tools and development and validation of novel clinical prediction models for pulmonary tuberculosis screening among outpatients living with HIV: an individual participant data meta-analysis

Ashar Dhana[1], Rishi K. Gupta [2], Yohhei Hamada [2,3], Andre P. Kengne[4], Andrew D. Kerkhoff [5], Christina Yoon[6], Adithya Cattamanchi[6], Byron W.P. Reeve[7], Grant Theron [7], Gcobisa Ndlangalavu[7], Robin Wood [8], Paul K. Drain [9], Claire J. Calderwood[2,10,11], Mahdad Noursadeghi[12], Tom Boyles[13], Graeme Meintjes[1,14], Gary Maartens[1,14] and David A. Barr[15]

[1]Department of Medicine, University of Cape Town, Cape Town, South Africa. [2]Institute for Global Health, University College London, London, UK. [3]Centre for International Cooperation and Global TB Information, The Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Tokyo, Japan. [4]Non-communicable Diseases Research Unit, South African Medical Research Council, Cape Town, South Africa. [5]Division of HIV, Infectious Diseases and Global Medicine, Zuckerberg San Francisco General Hospital and Trauma Center, University of California San Francisco, San Francisco, CA, USA. [6]Department of Medicine, Division of Pulmonary and Critical Care Medicine and Center for Tuberculosis, University of California San Francisco, San Francisco, CA, USA. [7]DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa. [8]Institute of Infectious Disease and Molecular Medicine (IDM), Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. [9]Departments of Global Health, Medicine, and Epidemiology, University of Washington, Seattle, WA, USA. [10]Clinical Research Department, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [11]The Research Unit Zimbabwe, Biomedical Research and Training Institute, Harare, Zimbabwe. [12]Division of Infection and Immunity, University College London, London, UK. [13]Helen Joseph Hospital, Johannesburg, South Africa. [14]Wellcome Centre for Infectious Diseases Research in Africa (CIDRI-Africa), Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa. [15]Institute of Infection and Global Health, University of Liverpool, Liverpool, UK.

Corresponding author: Ashar Dhana (ashardhana@live.com)

Abstract

Background: The World Health Organization (WHO) recommends that outpatient people living with HIV (PLHIV) undergo tuberculosis screening with the WHO four-symptom screen (W4SS) or C-reactive protein (CRP) (5 mg·L^{-1} cut-off) followed by confirmatory testing if screen positive. We conducted an individual participant data meta-analysis to determine the performance of WHO-recommended screening tools and two newly developed clinical prediction models (CPMs).

Methods: Following a systematic review, we identified studies that recruited adult outpatient PLHIV irrespective of tuberculosis signs and symptoms or with a positive W4SS, evaluated CRP and collected sputum for culture. We used logistic regression to develop an extended CPM (which included CRP and other predictors) and a CRP-only CPM. We used internal–external cross-validation to evaluate performance.

Results: We pooled data from eight cohorts (n=4315 participants). The extended CPM had excellent discrimination (C-statistic 0.81); the CRP-only CPM had similar discrimination. The C-statistics for WHO-recommended tools were lower. Both CPMs had equivalent or higher net benefit compared with the WHO-recommended tools. Compared with both CPMs, CRP (5 mg·L^{-1} cut-off) had equivalent net benefit across a clinically useful range of threshold probabilities, while the W4SS had a lower net benefit. The W4SS would capture 91% of tuberculosis cases and require confirmatory testing for 78% of participants.

CRP (5 mg·L^{-1} cut-off), the extended CPM (4.2% threshold) and the CRP-only CPM (3.6% threshold) would capture similar percentages of cases but reduce confirmatory tests required by 24, 27 and 36%, respectively. *Conclusions:* CRP sets the standard for tuberculosis screening among outpatient PLHIV. The choice between using CRP at 5 mg·L^{-1} cut-off or in a CPM depends on available resources.

Background

In 2021, there were 187 000 tuberculosis deaths among people living with HIV (PLHIV) [1]. Approximately half of HIV-associated tuberculosis deaths go undiagnosed [2] and appropriate testing and treatment may avert these undiagnosed deaths. Confirmatory testing (*e.g.* Xpert MTB/RIF Ultra (Xpert Ultra)) for all PLHIV is often not feasible in low-resource settings, meaning that screening strategies are needed to determine who needs further confirmatory testing.

According to the World Health Organization (WHO), a tuberculosis screening tool should meet optimal (95% sensitivity, 80% specificity) or minimum (90% sensitivity, 70% specificity) performance characteristics [3]. Since 2011, the WHO has recommended that outpatient PLHIV be screened for tuberculosis with the WHO four-symptom screen (W4SS) (comprising any one of current cough, fever, night sweats or weight loss) [4], followed by confirmatory testing (*e.g.* Xpert Ultra) if the screen is positive. However, the specificity of the W4SS is low in some subgroups (*e.g.* antiretroviral therapy (ART)-naïve PLHIV who still represent a third of all PLHIV) [5], resulting in large numbers of unnecessary, expensive confirmatory testing. Recently, the WHO also recommended chest radiography and C-reactive protein (CRP) as a screening tool [6, 7]. Chest radiography alone has suboptimal diagnostic performance [6]. CRP (5 mg·L^{-1} cut-off) showed similar sensitivity but higher specificity than the W4SS [5]. CRP can be done using a point-of-care assay at ~$2 with results in <3 min. The W4SS and CRP were recommended based on sensitivity and specificity, but their clinical utility is unknown.

Clinical prediction models (CPMs), which combine multiple predictors, may also be used for screening. Although there are a few CPMs for tuberculosis screening in PLHIV [8–15], these CPMs have limitations. Some have been developed using many predictors relative to number of events [8–10, 14] or categorised continuous variables [8–11]. Some have also shown suboptimal performance at external validation [8, 9], not undergone extensive external validation [9–11, 14, 15] or not been assessed for clinical usefulness [8–10, 14, 15].

Thus, we sought to overcome the limited data on clinical utility of WHO-recommended tuberculosis screening tools and methodological limitations in the development and validation of current CPMs for tuberculosis screening. Using data from eight cohorts identified following a systematic review [5], we performed an individual participant data (IPD) meta-analysis to 1) develop and validate CPMs that incorporate CRP for active pulmonary tuberculosis among outpatient PLHIV and 2) compare the performance and clinical utility of WHO-recommended screening tools to the newly developed CPMs.

Methods

We reported our findings according to the TRIPOD and PRISMA-IPD statements [16, 17]. We also adhered to additional guidelines for developing and validating CPMs in an IPD meta-analysis [18–20].

Data sources and study population

We previously conducted a systematic review to compare the accuracy of several tuberculosis screening tools with the W4SS in outpatient PLHIV [5]. The search of the systematic review was conducted from 1 January 2011 (the year the W4SS was first recommended by the WHO) to 12 March 2018. From that systematic review, we identified and obtained IPD for prospective cross-sectional studies, observational studies and randomised trials conducted in facility-based, active case-finding settings that systematically measured CRP and collected sputum for culture from outpatient PLHIV regardless of signs and symptoms of tuberculosis. We only included studies that measured CRP since we aimed to assess if a multivariable modelling approach could improve CRP performance. From the systematic review, we identified four studies (five cohorts) from active case-finding settings [21–24], comprising outpatient PLHIV not on ART (figure S1). Active case finding involves systematic screening of PLHIV irrespective of symptoms. Passive case finding involves patients recognising and seeking care for their symptoms (often defined in research studies as a positive W4SS). After discussion with policy makers and domain experts, we supplemented the active case-finding IPD identified through systematic review with two further studies (three cohorts) from passive case-finding settings [15, 25], comprising outpatient PLHIV receiving and not receiving ART. The purpose of this protocol change was to assess the effect of case-finding status on model performance, because case-finding status was deemed to be of significant interest to policy makers.

Defining the outcome of prediction models

The primary outcome was active tuberculosis, defined as positive culture of *Mycobacterium tuberculosis* complex from sputum.

Candidate predictors and sample size

Since all included studies evaluated CRP, we assessed these studies for availability of several variables considered *a priori* for potential inclusion in the CPMs. We selected variables following expert clinical experience and a systematic review [5] that were assessable immediately and that are readily available in resource-limited settings. The clinical variables we evaluated were age, sex, W4SS components (cough, fever, night sweats and weight loss) and body mass index (BMI). The laboratory variables we evaluated were CD4 count and CRP. The study-level variable we evaluated was case-finding setting (active *versus* passive case finding).

The derived population was considered sufficient [26] and sample size calculations are provided in the appendix.

Statistical analysis

We developed two CPMs, as follows: an extended CPM, which considered all candidate predictors; and a CRP-only CPM, which only included CRP as a continuous predictor along with spline transformations. The CRP-only CPM differs from CRP (5 mg·L^{-1} cut-off), which is recommended by the WHO and is a typical binary diagnostic test. Figure 1 summarises the steps taken to develop and validate each CPM.

CPM development

We performed single-level multiple imputation within each cohort to deal with missing data (detailed in the appendix). We created 10 imputed datasets. All further analyses were performed in each of the imputed datasets and pooled using Rubin's rules [27].

We used a logistic regression approach for variable selection and CPM development with active tuberculosis as a binary outcome. To model continuous variables, we used restricted cubic splines with a default of four knots. For the extended CPM, we performed backward stepwise selection with bootstrapping to select the most predictive variables using the Akaike information criterion [28]. We kept variables retained in ⩾70% of bootstrap samples and ⩾5 of 10 multiply imputed datasets. This process led
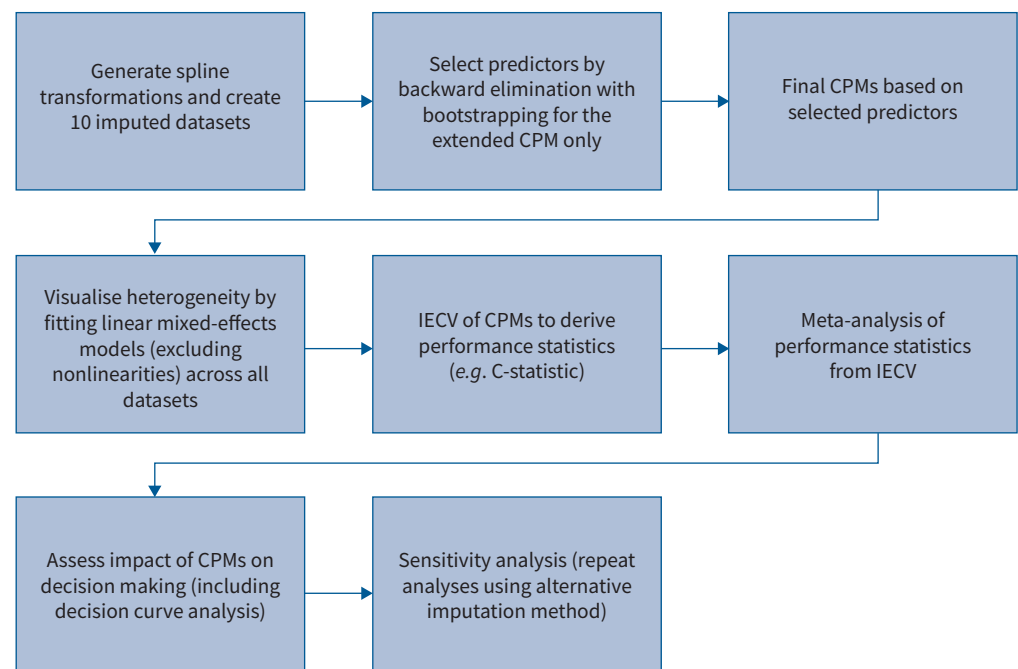

FIGURE 1 Steps taken to develop and validate each clinical prediction model (CPM). IECV: internal–external cross-validation.

to a final CPM based on the selected predictors along with their corresponding estimated β coefficients and the associated intercept term. To visualise predictor heterogeneity, we fitted linear mixed-effects models across all datasets (excluding nonlinearities).

Internal–external cross-validation (IECV)

To assess CPM generalisability, we used IECV, which involved several steps [18]. First, the CPM is developed in all but one study. Second, the omitted study is used to externally validate the CPM and derive performance statistics. Third, this process is repeated until each study has a chance to be omitted. We performed IECV on each imputed dataset and pooled performance statistics using Rubin's rules. In the omitted study, we also compared both CPMs to WHO-recommended screening tools and one other published CPM [9]. Other published CPMs were not evaluated because some predictors were not measured in some or all cohorts [8, 10–15].

We calculated several performance statistics. We assessed discrimination, which quantifies how well a CPM can differentiate between those that have tuberculosis and those that do not, using the C-statistic. We considered a C-statistic of $\geqslant 0.7$ and $\geqslant 0.8$ as acceptable and excellent performance, respectively [29]. We then assessed calibration, which refers to agreement between expected and observed outcomes, using calibration in the large (value of 0 indicates perfect calibration), calibration slope (value of 1 indicates perfect calibration) and calibration plots.

We performed a univariate random-effects meta-analysis of performance statistics derived from the IECV [30, 31]. To assess heterogeneity, we visually examined forest plots. We also performed a multivariate meta-analysis to pool the C-statistic and calibration slope [31]. We calculated the joint probability that the CPMs would achieve a C-statistic of >0.70, >0.75 or 0.80 and a calibration slope between 0.8 and 1.2 in future patients using bootstrapping.

We assessed clinical utility using two approaches. First, during IECV, we performed decision curve analyses by pooling (stacking) multiply imputed validation datasets [32, 33]. Decision curves show the net benefit of screening tools over a range of clinically relevant threshold probabilities. A threshold probability is the minimum probability of disease at which further diagnostic workup would be justified [33]. Net benefit is the difference between proportion of true positives and proportion of false positives weighted by the threshold probability. We chose a threshold probability range from 0 to 20% because it is unlikely that more than 20% risk would be required before confirmatory testing is recommended [34]. The CPMs were compared with a confirmatory testing for all strategy (*i.e.* sputum culture for all), confirmatory testing for none strategy (*i.e.* sputum culture for none), WHO-recommended screening tests (W4SS and CRP (5 mg·L^{-1} cut-off)) and an existing CPM [9]. Second, we assessed the trade-off between percentage of tuberculosis cases captured and percentage of participants needing confirmatory testing by applying WHO-recommended screening tests to the stacked validation cohorts during IECV and then comparing them to the newly developed CPMs at a threshold that provides similar sensitivity and 95% sensitivity (WHO optimal sensitivity requirements) [3].

In sensitivity analyses, we used an alternative imputation procedure using the aregImpute function in the rms package in R. We performed all analyses using *pmsampsize, mice, rms, metamisc, meta, metafor, lme, mada* and *dcurves* packages in R (version 3.6.1).

Results

Study population

IPD were provided for all six eligible studies (eight cohorts). The cohorts collected data between 2010 and 2020 (table S1). Six cohorts were from South Africa. Five were active case-finding cohorts and three were passive case-finding cohorts. Table 1 and table S2 show participant characteristics overall and by study, respectively. We included 4315 participants. Of 4209 participants with sputum culture results, 652 (15%) had tuberculosis. The prevalence of tuberculosis in active case-finding cohorts was 13% and in passive case-finding cohorts was 29%. Most participants (85%) were recruited from active case-finding settings. Most (91%) participants were not on ART. Table S3 shows missing data by study.

CPM development

After backward stepwise selection on the full dataset, 5 of 10 candidate predictors were selected in $\geqslant 70\%$ of bootstrap samples in $\geqslant 5$ of 10 multiply imputed datasets (table S4). The predictors selected were age (60% of multiply imputed datasets), BMI (100%), CD4 count (100%), CRP (100%) and cough (50%). The predictors not selected were sex (0%), fever (0%), weight loss (0%), night sweats (0%) and case finding (0%). Table S5 shows all coefficients and knot locations of the CPMs. CRP had the strongest

TABLE 1 Summary of main characteristics for all participants		
Variable	Overall[#]	N[¶]
Participants	4315 (100)	
Demographics		
Active case-finding	3667 (85)	4315
Age (years)	33.2 (27–40)	4315
Female	2381 (55)	4315
HIV history		
On ART	380 (9)	4315
CD4 count (cells·μL^{-1})	204 (93–319)	4188
CD4 ⩽200 cells·μL^{-1}	2056 (49)	4188
Clinical characteristics		
History of tuberculosis	602 (14)	4309
Cough	2256 (52)	4314
Fever	1541 (36)	4283
Weight loss	2638 (62)	4283
Night sweats	1674 (39)	4313
Cough ⩾2 weeks	1455 (34)	4306
Tuberculosis diagnostic tests		
Sputum culture+	652 (15)	4209
Laboratory tests		
BMI (kg·m^{-2})	22 (19.4–25.8)	4306
CRP (mg·L^{-1})	6.4 (2.5–38.6)	4093
Hb (g·dL^{-1})	12.5 (11–13.9)	2453

ART: antiretroviral therapy; BMI: body mass index; CRP: C-reactive protein; Hb: haemoglobin. [#]: Data are count (%) or median (25–75th percentiles). [¶]: Participants with data available for variable.

association with the outcome from all predictors (figure S2). Figure S3 shows the associations (excluding cubic spline transformations) between each retained predictor and the outcome after fitting a linear mixed-effects model across datasets.

CPM validation

For the extended CPM, figure 2 shows forest plots for performance statistics calculated during IECV. Discrimination was excellent. The pooled C-statistic was 0.81 (0.77, 0.84). C-statistics were consistent within active case-finding settings but heterogenous in passive case-finding settings. Calibration was adequate. Pooled calibration in the large was 0.02 (−0.15–0.20; value of 0 indicates perfect calibration). There was slight underestimation of risk in the Lawn and Scottsdene cohorts. Pooled calibration slope was 0.96 (0.83–1.09; value of 1 indicates perfect calibration). The calibration plots suggest reasonable agreement between predicted and observed risk for most cohorts (figure S4) but suboptimal calibration in the Reeve cohort (figure S4A). The joint probability of achieving a C-statistic >0.80 and a calibration slope between 0.8 and 1.2 was 54% (table S6).

For the CRP-only CPM, figure 3 shows forest plots for performance statistics calculated during IECV. Discrimination was similar to that of the extended CPM with a pooled C-statistic of 0.79 (0.74–0.83); calibration was also adequate. Pooled calibration in the large was 0.01 (−0.20–0.22). There was slight underestimation of risk in the Scottsdene cohort and slight overestimation in the Reeve cohort. Pooled calibration slope was 0.97 (0.86–1.08). Figure S5 shows that calibration plots were similar to those of the extended CPM with suboptimal calibration in the Reeve and Scottsdene cohorts (figures S5A and S5G). The joint probability of achieving a C-statistic >0.80 and a calibration slope between 0.8 and 1.2 was 38% (table S6).

Table 2 compares performance statistics for both CPMs with WHO-recommended screening tools and another published CPM. Both CPMs had higher discrimination.

Assessment of clinical utility

Figure 4 and figure S6 show decision curve analyses from the pooled IECV validation sets. The net benefit of both CPMs was equivalent or higher than that of other strategies across the range of threshold probabilities. Between-threshold probabilities of 0 and 3.1%, net benefit of a "confirmatory testing for all" strategy was at least equivalent to that of both CPMs. Between-threshold probabilities of 3.1 and 7.7%, net
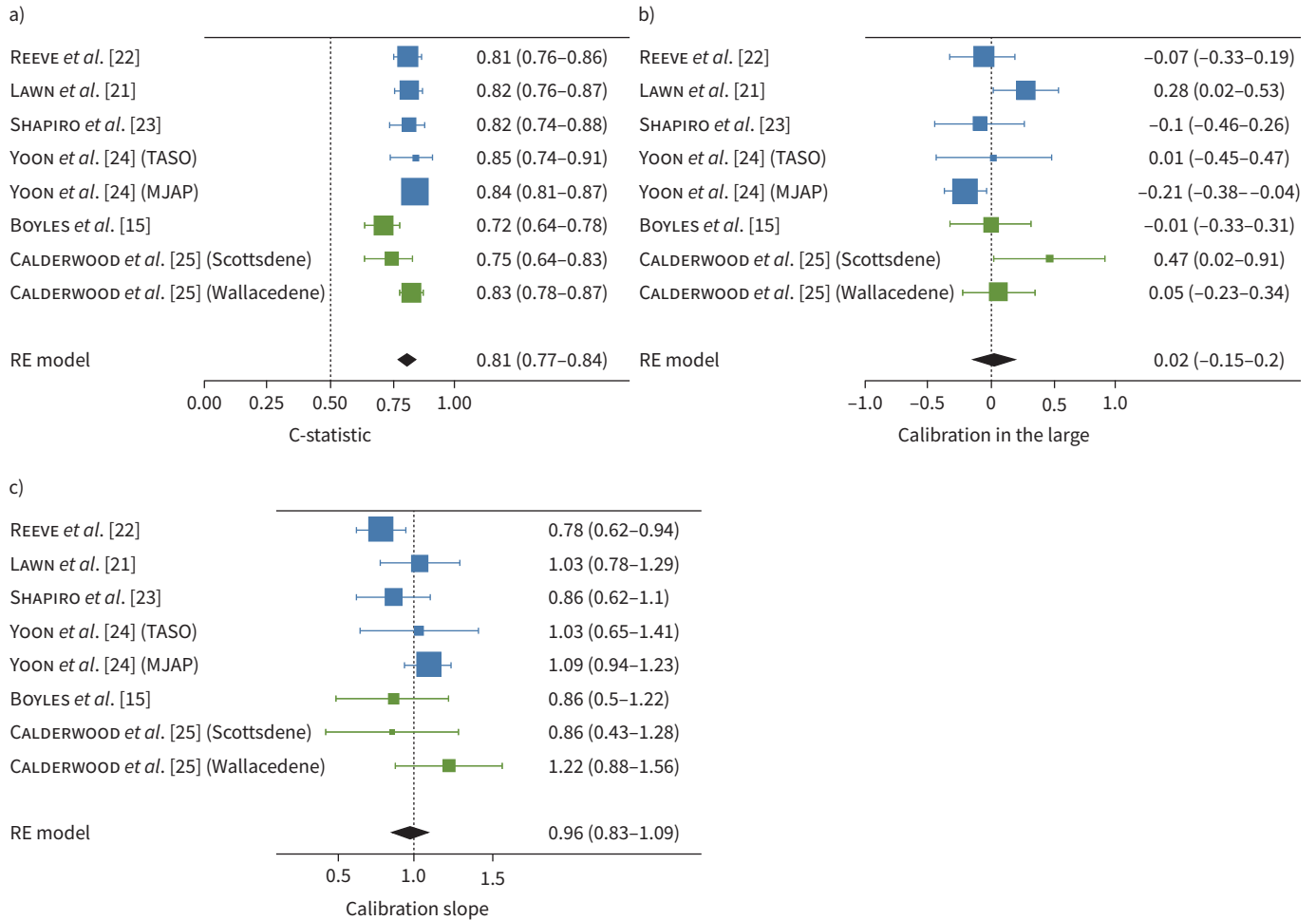
a)

Reeve et al. [22]		0.81 (0.76–0.86)
Lawn et al. [21]		0.82 (0.76–0.87)
Shapiro et al. [23]		0.82 (0.74–0.88)
Yoon et al. [24] (TASO)		0.85 (0.74–0.91)
Yoon et al. [24] (MJAP)		0.84 (0.81–0.87)
Boyles et al. [15]		0.72 (0.64–0.78)
Calderwood et al. [25] (Scottsdene)		0.75 (0.64–0.83)
Calderwood et al. [25] (Wallacedene)		0.83 (0.78–0.87)
RE model		0.81 (0.77–0.84)

C-statistic (x-axis: 0.00 0.25 0.50 0.75 1.00)

b)

Reeve et al. [22]		−0.07 (−0.33–0.19)
Lawn et al. [21]		0.28 (0.02–0.53)
Shapiro et al. [23]		−0.1 (−0.46–0.26)
Yoon et al. [24] (TASO)		0.01 (−0.45–0.47)
Yoon et al. [24] (MJAP)		−0.21 (−0.38–−0.04)
Boyles et al. [15]		−0.01 (−0.33–0.31)
Calderwood et al. [25] (Scottsdene)		0.47 (0.02–0.91)
Calderwood et al. [25] (Wallacedene)		0.05 (−0.23–0.34)
RE model		0.02 (−0.15–0.2)

Calibration in the large (x-axis: −1.0 −0.5 0 0.5 1.0)

c)

Reeve et al. [22]		0.78 (0.62–0.94)
Lawn et al. [21]		1.03 (0.78–1.29)
Shapiro et al. [23]		0.86 (0.62–1.1)
Yoon et al. [24] (TASO)		1.03 (0.65–1.41)
Yoon et al. [24] (MJAP)		1.09 (0.94–1.23)
Boyles et al. [15]		0.86 (0.5–1.22)
Calderwood et al. [25] (Scottsdene)		0.86 (0.43–1.28)
Calderwood et al. [25] (Wallacedene)		1.22 (0.88–1.56)
RE model		0.96 (0.83–1.09)

Calibration slope (x-axis: 0.5 1.0 1.5)

FIGURE 2 Forest plots showing a) extended clinical prediction model discrimination, b) calibration in the large and c) calibration slope. MJAP: Makerere University Joint AIDS Program; RE: random effects; TASO: The AIDS Support Organization. Studies in blue are from active case-finding settings and studies in green are from passive case-finding settings.

benefit of CRP (5 mg·L^{-1} cut-off) was at least equivalent to that of both CPMs. At a threshold probability >7.7%, the net benefit of both CPMs was higher than that of the other strategies. The net benefit of one published CPM [9] was generally lower than that of both CPMs except at very low threshold probabilities (<3.1%). Results were similar when excluding passive case-finding cohorts (figure S6). The net benefit of the W4SS was lower than that of both CPMs and CRP (5 mg·L^{-1} cut-off) across all threshold probabilities.

We applied CRP (5 mg·L^{-1} cut-off) to the stacked multiply imputed datasets (table 3). CRP (5 mg·L^{-1}) would have captured 91% of tuberculosis cases and resulted in confirmatory testing for 54% of participants. In comparison, to capture a similar percentage of cases, both CPMs would have resulted in confirmatory testing for a similar percentage of participants. To capture 95% of those with tuberculosis, the extended and CRP-only CPMs would have resulted in confirmatory testing for 74 and 75% of participants, respectively. Excluding passive case-finding cohorts, since all participants in those cohorts were W4SS positive, we applied the W4SS to the stacked multiply imputed datasets (table 3). The W4SS would have captured 91% of tuberculosis cases and resulted in confirmatory testing for 78% of participants. CRP (5 mg·L^{-1} cut-off), the extended CPM (4.2% threshold) and the CRP-only CPM (3.6% threshold) would have captured a similar percentage of cases but required confirmatory tests for only 50, 59 and 57% of participants, respectively.

Sensitivity analyses
We repeated analyses for the extended CPM using an alternative imputation method (figure S7 and table S7). The results were similar to the main analyses.
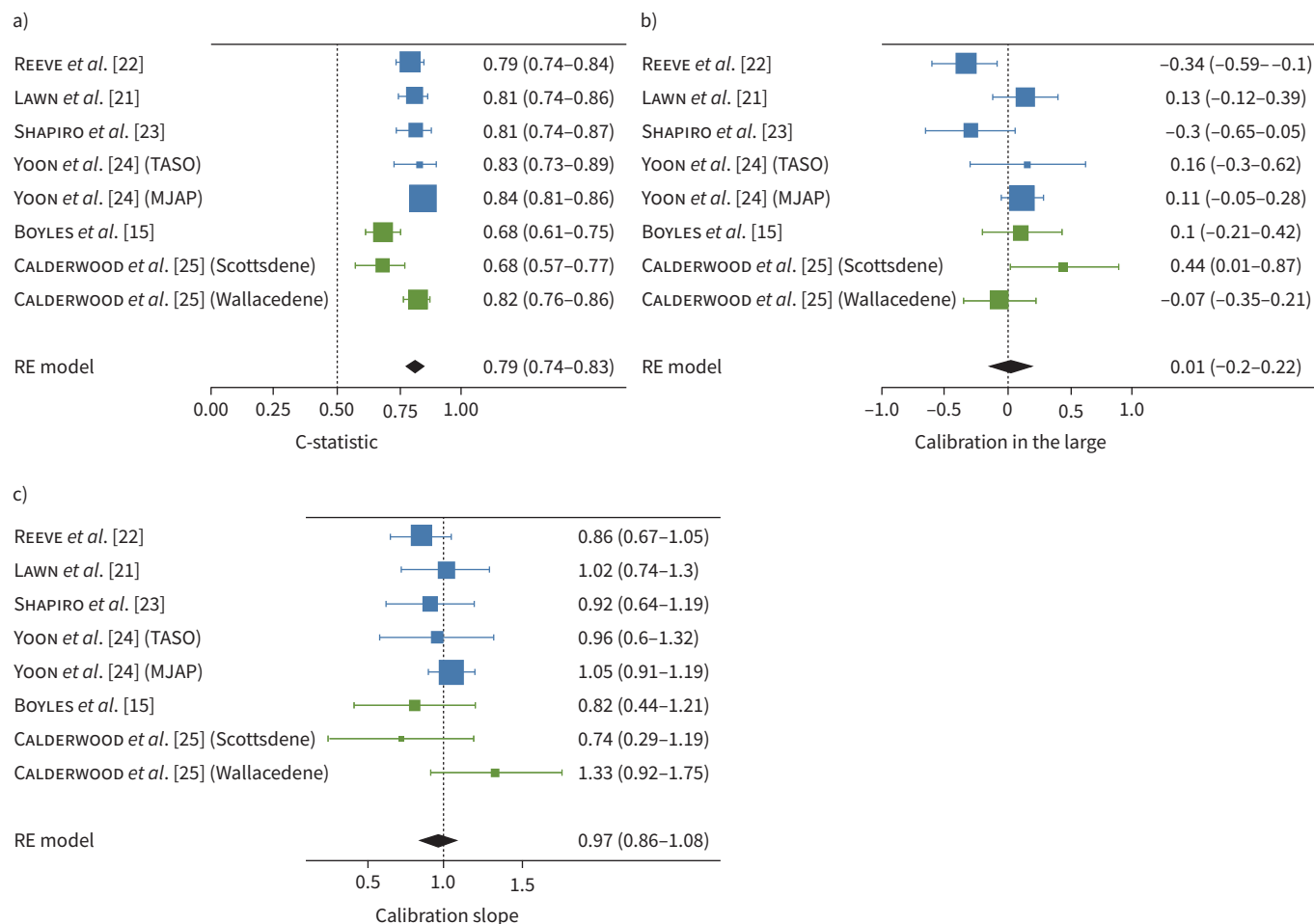
FIGURE 3 Forest plots showing a) C-reactive protein only clinical prediction model discrimination, b) calibration in the large and c) calibration slope. MJAP: Makerere University Joint AIDS Program; RE: random effects; TASO: The AIDS Support Organization. Studies in blue are from active case-finding settings and studies in green are from passive case-finding settings.

Discussion

We investigated the utility of WHO-recommended screening tools (W4SS and CRP (5 mg·L^{-1} cut-off)) and developed and validated two CPMs for pulmonary tuberculosis screening in outpatient PLHIV using eight cohorts (4315 participants). At validation, the extended CPM showed excellent discrimination and

TABLE 2 Performance statistics of extended and C-reactive protein (CRP)-only clinical prediction models (CPMs), World Health Organization (WHO)-recommended tools and other CPMs[#]

Model/test	Concordance statistic	Calibration in the large	Calibration slope
Extended CPM	0.81 (0.77–0.84)	0.02 (−0.15–0.20)	0.96 (0.83–1.09)
CRP-only CPM	0.79 (0.74–0.83)	0.01 (−0.20–0.22)	0.97 (0.86–1.08)
W4SS[¶]	0.57 (0.51–0.63)		
CRP (5 mg·dL^{-1})[+]	0.70 (0.63–0.75)		
CPM from HANIFA et al. [9]	0.71 (0.68–0.75)	0.41 (−0.21–1.03)	1.05 (0.79–1.32)

[#]: The extended CPM contained CRP, age, body mass index, CD4 cell count and cough as predictors along with spline transformations for continuous variables. The CRP-only CPM only included CRP as a predictor along with spline transformations. [¶]: WHO four-symptom screen (W4SS) performance statistics calculated only for active case-finding datasets as all participants in passive case-finding datasets were W4SS positive. [+]: CRP (5 mg·dL^{-1} cut-off) performance statistics exclude the study by BOYLES et al. [15] as all participants in that study had a CRP level ⩾5 mg·L^{-1}.
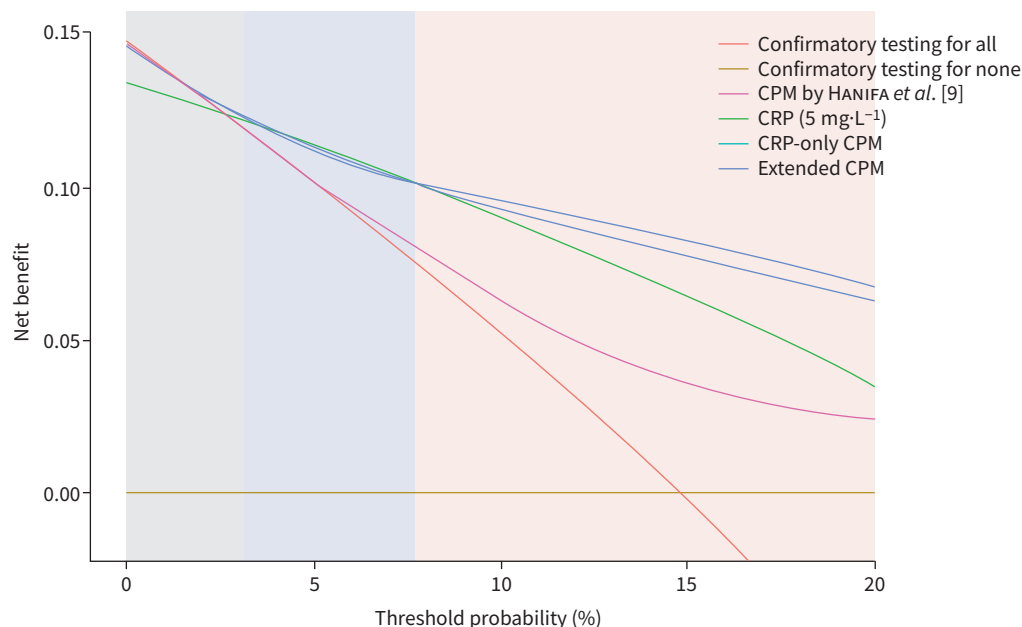
FIGURE 4 Decision curve analysis comparing the extended and C-reactive protein (CRP)-only clinical prediction models (CPMs) to other tools or strategies other than the World Health Organization four-symptom screen among active and passive case-finding cohorts. Grey shaded region: if testing 32 or more patients per tuberculosis (TB) case diagnosed, confirmatory testing for all has at least an equivalent net benefit to that of the extended and CRP-only models. Blue shaded region: if testing between 32 and 13 patients per TB case diagnosed, CRP (5 mg·L^{-1} cut-off) has at least an equivalent net benefit to that of the extended and the CRP-only models. Red shaded region: if testing 13 or fewer patients per TB case diagnosed, the extended model has more net benefit than the CRP-only model. The results exclude the study by BOYLES et al. [15], as all participants in that study a CRP level ⩾5 mg·L^{-1}.

TABLE 3 Trade-off between percentage of tuberculosis cases captured and percentage of participants needing confirmatory testing for the extended clinical prediction model (CPM), C-reactive protein (CRP)-only CPM and World Health Organization (WHO)-recommended tools using the 10 stacked multiply imputed datasets[#]

CPM or tool	CPM-based tuberculosis threshold (%)	Percentage of tuberculosis cases captured	Percentage of all needing confirmatory testing	Number of confirmatory tests to capture one tuberculosis case
Active and passive case-finding cohorts (n=41 080)[#,¶]				
CRP 5 mg·L^{-1}		91	54	4
Extended CPM[¶]	4.9	91	56	4.1
CRP-only CPM[¶]	4.2	91	54	4
Extended CPM[+]	2.9	95	74	5.3
CRP-only CPM[+]	3.0	95	75	5.3
Active case-finding cohorts (n=36 670)[f]				
W4SS		91	78	6.5
CRP 5 mg·L^{-1}		89	50	4.3
Extended CPM[§]	4.2	91	59	4.9
CRP-only CPM[§]	3.6	91	57	4.8
Extended CPM[+]	2.7	95	76	6.1
CRP-only CPM[+]	2.8	95	80	6.4

[#]: Excludes study by BOYLES et al. [15], as all participants in that study had CRP level ⩾5 mg·L^{-1}. [¶]: For both CPMs, thresholds were selected to capture a similar percentage of tuberculosis cases compared with CRP at 5 mg·L^{-1} cut-off (91%). [+]: For both CPMs, thresholds were selected to capture a similar percentage of tuberculosis cases compared with an ideal triage test according to WHO target product profile (95%). [§]: For both CPMs, thresholds were selected to capture a similar percentage of tuberculosis cases compared with the WHO four-symptom screen (W4SS). [f]: Excludes three passive case-finding cohorts, as all participants in those studies had a positive W4SS.

adequate calibration; the CRP-only CPM showed similar performance. The W4SS and CRP ($5\ \mathrm{mg \cdot L^{-1}}$ cut-off) had lower discrimination. Both CPMs had equivalent or higher net benefit across all threshold probabilities compared with other tools or strategies. However, CRP ($5\ \mathrm{mg \cdot L^{-1}}$ cut-off) demonstrated similar net benefit to both CPMs over a clinically plausible range of threshold probabilities; CRP ($5\ \mathrm{mg \cdot L^{-1}}$ cut-off) also met WHO minimum sensitivity requirements (90% sensitivity). At lower threshold probabilities or if WHO optimal sensitivity requirements (95% sensitivity) are preferred, both CPMs and a "confirmatory testing for all" strategy had similar net benefit. The W4SS had suboptimal net benefit. CRP ($5\ \mathrm{mg \cdot L^{-1}}$ cut-off) had similar sensitivity to the W4SS but required 36% fewer confirmatory tests.

By assessing clinical utility, we provide further evidence of CRP's value for tuberculosis screening in outpatient PLHIV. In a recent meta-analysis, CRP ($5\ \mathrm{mg \cdot L^{-1}}$ cut-off) showed similar sensitivity but higher specificity compared with the W4SS [5], leading to its inclusion in updated WHO tuberculosis screening guidelines [6]. It is recommended that emerging biomarkers be evaluated against available tools [35]. Our findings suggest that CRP and the newly developed CPMs be used as a benchmark to evaluate emerging biomarkers for tuberculosis screening. CRP may also be combined with other biomarkers to improve predictive performance. The addition of clinical characteristics (*i.e.* W4SS symptoms) to CRP provided minimal extra information, since both CPMs showed similar performance. The W4SS is a key component of tuberculosis screening guidelines but has suboptimal utility. Variable selection further demonstrated the limited role of symptoms in predicting tuberculosis as only one of the four W4SS symptoms was retained during backward selection.

Although CRP ($5\ \mathrm{mg \cdot L^{-1}}$ cut-off) and both CPMs had high net benefit across a wide range of thresholds, a "confirmatory testing for all" strategy may be considered following economic analyses and if a setting has resources to perform many confirmatory tests per case diagnosed. Given the high prior probability of tuberculosis in this study (tuberculosis prevalence between 25–38% in passive case-finding cohorts and 10–17% in active case-finding cohorts not yet on ART), a "confirmatory testing for all" strategy may be plausible.

We externally validated a published CPM by Hanifa *et al.* [9], which showed suboptimal utility and performance compared with our CPMs. Hanifa *et al.* [9]. included similar predictors but did not include CRP or account for nonlinear associations. Auld *et al.* [8] recently developed a CPM for tuberculosis in outpatient PLHIV and validated the CPM in three cohorts. The CPM included W4SS symptoms, sex, smoking status, temperature, BMI and haemoglobin as predictors. However, at a cut-off that provides similar sensitivity to the W4SS, the score did not improve specificity. The score was also externally validated using a cohort included in this article, showing much lower discrimination than the extended CPM (C-statistic of 0.63 *versus* 0.82 for the extended CPM) [21]. Baik *et al.* [11] recently developed a CPM for tuberculosis in symptomatic outpatients irrespective of HIV status. However, performance was not assessed in PLHIV. Balcha *et al.* [10] developed a relatively complex CPM for tuberculosis among outpatients with a positive W4SS. However, the CPM has not been validated internally or externally. Similarly, the TBscore has been developed but is complex, consisting of 11 symptoms and signs, and has low specificity (36%) [12].

Our study has several strengths. This study is the only one to validate a CPM and other tools for tuberculosis using the recommended IECV framework [18]. We included a large population of outpatient PLHIV from eight different settings to evaluate generalisability. We used various measures of clinical utility, including net benefit and the trade-off between number of tuberculosis cases captured and unnecessary additional confirmatory testing. For CPM development, we used multiple imputation to handle missing data, selected readily available predictors, avoided categorisation of continuous variables and accounted for nonlinear relationships. We have also presented the full equation for both CPMs to enable use in clinical practice. In the future, the CPMs may be easily presented using a website calculator, mobile app or spreadsheet software. Finally, we adhered to the TRIPOD statement and additional guidelines [18–20].

Our study has several limitations. First, active case-finding study populations did not include PLHIV on ART and passive case-finding cohorts only comprised 15% of all data. Therefore, results should be extrapolated with caution to these subpopulations. However, PLHIV not on ART – who comprised 91% of participants – currently still represent a third of all PLHIV (~13 million people) [36] and have a high tuberculosis prevalence (~10–15%) [5]. Second, all cohorts were drawn from high-burden outpatient settings in South Africa and Uganda, meaning results may not generalise to low-burden settings. Our results might also not be generalisable to children. Third, we did not include certain well-known predictors of tuberculosis, such as haemoglobin [37], because of missing data. We were also unable to evaluate chest radiography – another WHO-recommended screening tool – since only one study performed chest radiography. However, chest

radiography has suboptimal diagnostic performance as a screening tool and is only recommended in combination with the W4SS [6]. Besides, it is often unavailable in primary health centres [38]. Fourth, although our results are largely applicable to pulmonary tuberculosis, extrapulmonary tuberculosis is less likely in outpatient settings. Furthermore, our results are based on an imperfect reference standard since some pulmonary tuberculosis cases may have been culture negative. Fifth, we were unable to evaluate several published CPMs with predictors that were not measured in some or all cohorts [8, 10–15]. Finally, we did not investigate the cost and resource implications of CRP-based strategies. Further economic analyses (*e.g.* cost minimisation and budget impact analyses) are needed for these strategies.

In conclusion, our findings define optimal tuberculosis screening strategies in outpatient PLHIV based on currently available data, accounting for the trade-off between the number of tuberculosis cases diagnosed and number of confirmatory tests performed. CRP (5 mg·L^{-1} cut-off) – which has been recently recommended by WHO – showed optimal net benefit across a plausible range of thresholds. Two newly developed CPMs that incorporate CRP as a predictor may add value at more extreme threshold probabilities – where resources allow more or fewer confirmatory tests per diagnosed case. A "confirmatory testing for all" strategy might also be considered if resources permit. Conversely, the W4SS showed suboptimal utility. CRP (either alone or as part of a CPM) sets the standard for tuberculosis screening in outpatient PLHIV and the newly developed CPMs may also be used as a benchmark to evaluate future biomarkers or combined with other biomarkers to improve predictive performance.

Provenance: Submitted article, peer reviewed.

References

1   World Health Organization. Global tuberculosis report 2022. Geneva, World Health Organization, 2022.

2   Gupta RK, Lucas SB, Fielding KL, *et al.* 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: 1987–2002.

3   World Health Organization. High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28–29 April 2014, Geneva, Switzerland. Geneva, World Health Organization, 2014.

4   Getahun H, Kittikraisak W, Heilig CM, *et al.* Development of a standardized screening rule for tuberculosis in people living with HIV in resource-constrained settings: individual participant data meta-analysis of observational studies. *PLoS Med* 2011; 8: e1000391.

5   Dhana A, Hamada Y, Kengne AP, *et al.* Tuberculosis screening among ambulatory people living with HIV: a systematic review and individual participant data meta-analysis. *Lancet Infect Dis* 2022; 22: 507–518.

6   World Health Organization. WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. Geneva, World Health Organization, 2021. Report No.: 9240022678.

7   World Health Organization. Rapid communication on systematic screening for tuberculosis. Geneva, World Health Organization, 2020. Report No.: 9240016554.

8   Auld AF, Kerkhoff AD, Hanifa Y, *et al.* Derivation and external validation of a risk score for predicting HIV-associated tuberculosis to support case finding and preventive therapy scale-up: a cohort study. *PLoS Med* 2021; 18: e1003739.

9   Hanifa Y, Fielding KL, Chihota VN, *et al.* A clinical scoring system to prioritise investigation for tuberculosis among adults attending HIV clinics in South Africa. *PLoS One* 2017; 12: e0181519.

10   Balcha TT, Skogmar S, Sturegard E, *et al.* A clinical scoring algorithm for determination of the risk of tuberculosis in HIV-infected adults: a cohort study performed at Ethiopian health centers. *Open Forum Infect Dis* 2014; 1: ofu095.

11   Baik Y, Rickman HM, Hanrahan CF, *et al.* A clinical score for identifying active tuberculosis while awaiting microbiological results: development and validation of a multivariable prediction model in sub-Saharan Africa. *PLoS Med* 2020; 17: e1003420.

12   Aunsborg JW, Honge BL, Jespersen S, *et al.* A clinical score has utility in tuberculosis case-finding among patients with HIV: a feasibility study from Bissau. *Int J Infect Dis* 2020; 92S: S78–S84.

13   Rangaka MX, Wilkinson RJ, Glynn JR, *et al.* Effect of antiretroviral therapy on the diagnostic accuracy of symptom screening for intensified tuberculosis case finding in a South African HIV clinic. *Clin Infect Dis* 2012; 55: 1698–1706.

14   Nanta S, Kantipong P, Pathipvanich P, *et al.* Screening scheme development for active TB prediction among HIV-infected patients. *Southeast Asian J Trop Med Public Health* 2011; 42: 867–875.

15   Boyles TH, Nduna M, Pitsi T, *et al.* A clinical prediction score including trial of antibiotics and C-reactive protein to improve the diagnosis of tuberculosis in ambulatory people with HIV. *Open Forum Infect Dis* 2020; 7: ofz543.

16   Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Ann Intern Med* 2015; 162: 735–736.

17   Stewart LA, Clarke M, Rovers M, *et al.* Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD statement. *JAMA* 2015; 313: 1657–1665.

18   Debray TP, Riley RD, Rovers MM, *et al.* Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PLoS Med* 2015; 12: e1001886.

19   Ahmed I, Debray TP, Moons KG, *et al.* Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol* 2014; 14: 3.

20   Debray TP, Moons KG, Ahmed I, *et al.* A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013; 32: 3158–3180.

21   Lawn SD, Kerkhoff AD, Vogt M, *et al.* Diagnostic and prognostic value of serum C-reactive protein for screening for HIV-associated tuberculosis. *Int J Tuberc Lung Dis* 2013; 17: 636–643.

22   Reeve B, Ndlangalavu G, Palmer Z, *et al.* Accuracy of Xpert Ultra and Xpert MTB/RIF in people living with HIV initiating antiretroviral treatment who have minimal TB symptoms. 50th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease, Hyderabad, India, Oct 30–Nov 2, 2019; p. S115.

23   Shapiro AE, Hong T, Govere S, *et al.* C-reactive protein as a screening test for HIV-associated pulmonary tuberculosis prior to antiretroviral therapy in South Africa. *AIDS* 2018; 32: 1811–1820.

24   Yoon C, Semitala FC, Atuhumuza E, *et al.* Point-of-care C-reactive protein-based tuberculosis screening for people living with HIV: a diagnostic accuracy study. *Lancet Infect Dis* 2017; 17: 1285–1292.

25   Calderwood CJ, Reeve BWP, Mann T, *et al.* Clinical utility of C-reactive protein-based triage for presumptive pulmonary tuberculosis in South African adults. *J Infection* 2023; 86: 24–32.

26   Riley RD, Ensor J, Snell KIE, *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368: m441.

27   Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, John Wiley & Sons, 2004.

28   Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Biom J* 2018; 60: 431–449.

29   Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken, John Wiley & Sons, 2013.

30   Debray TP, Damen JA, Snell KI, *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017; 356: i6460.

31   Snell KI, Hua H, Debray TP, *et al.* Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2016; 69: 40–50.

32   Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26: 565–574.

33   Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; 3: 18.

34   Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; 352: i6.

35   Kerkhoff AD, Cattamanchi A, Muyoyeta M, *et al.* Validating novel diagnostic assays for tuberculosis in the context of existing tools. *Lancet Glob Health* 2021; 9: e1209.

36   Joint United Nations Programme on HIV/AIDS (UNAIDS). 2020 Global AIDS update - Seizing the Moment - Tackling entrenched inequalities to end epidemics, Geneva, UNAIDS, 2020.

37   Kerkhoff AD, Wood R, Vogt M, *et al.* Predictive value of anemia for tuberculosis in HIV-infected patients in Sub-Saharan Africa: an indication for routine microbiological investigation using new rapid assays. *J Acquir Immune Defic Syndr* 2014; 66: 33–40.

38   Huddart S, MacLean E, Pai M. Location, location, location: tuberculosis services in highest burden countries. *Lancet Glob Health* 2016; 4: e907–e908.