BMJ Open

Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults

Lauren S. Peetluk ,[1] Felipe M. Ridolfi,[2] Peter F. Rebeiro,[1,3] Dandan Liu,[4] Valeria C Rolla,[2] Timothy R. Sterling[3]

To cite: Peetluk LS, Ridolfi FM, Rebeiro PF, *et al*. Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults. *BMJ Open* 2021;**11**:e044687. doi:10.1136/bmjopen-2020-044687

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-044687).

Received 10 September 2020
Revised 09 February 2021
Accepted 17 February 2021

[1]Division of Epidemiology, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[2]Instituto Nacional de Infectologia Evandro Chagas, Rio de Janeiro, Brazil
[3]Division of Infectious Diseases, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[4]Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence to
Lauren S. Peetluk;
lauren.s.peetluk@vanderbilt.edu

ABSTRACT

Objective To systematically review and critically evaluate prediction models developed to predict tuberculosis (TB) treatment outcomes among adults with pulmonary TB.

Design Systematic review.

Data sources PubMed, Embase, Web of Science and Google Scholar were searched for studies published from 1 January 1995 to 9 January 2020.

Study selection and data extraction Studies that developed a model to predict pulmonary TB treatment outcomes were included. Study screening, data extraction and quality assessment were conducted independently by two reviewers. Study quality was evaluated using the Prediction model Risk Of Bias Assessment Tool. Data were synthesised with narrative review and in tables and figures.

Results 14 739 articles were identified, 536 underwent full-text review and 33 studies presenting 37 prediction models were included. Model outcomes included death (n=16, 43%), treatment failure (n=6, 16%), default (n=6, 16%) or a composite outcome (n=9, 25%). Most models (n=30, 81%) measured discrimination (median c-statistic=0.75; IQR: 0.68–0.84), and 17 (46%) reported calibration, often the Hosmer-Lemeshow test (n=13). Nineteen (51%) models were internally validated, and six (16%) were externally validated. Eighteen (54%) studies mentioned missing data, and of those, half (n=9) used complete case analysis. The most common predictors included age, sex, extrapulmonary TB, body mass index, chest X-ray results, previous TB and HIV. Risk of bias varied across studies, but all studies had high risk of bias in their analysis.

Conclusions TB outcome prediction models are heterogeneous with disparate outcome definitions, predictors and methodology. We do not recommend applying any in clinical settings without external validation, and encourage future researchers adhere to guidelines for developing and reporting of prediction models.

Trial registration The study was registered on the international prospective register of systematic reviews PROSPERO (CRD42020155782)

Strengths and limitations of this study

- ► Prediction models for tuberculosis treatment outcomes have the potential to inform interventions or treatment management protocols to promote cure among patients with tuberculosis at the greatest risk of unsuccessful treatment outcomes, but the methods and clinical utility of existing models had not been formally evaluated.
- ► This was the first systematic review of prediction models for tuberculosis treatment outcomes.
- ► The review used a comprehensive search strategy, conducted thorough bias assessment with the Prediction Model Risk of Bias Assessment Tool (PROBAST) tool, and offers recommendations for future model development and validation studies for predicting tuberculosis treatment outcomes.
- ► Evidence synthesis and quality assessment were limited by incomplete reporting in primary studies, as well as heterogeneities in study populations, such as multidrug resistance and age.
- ► External validation studies or studies written in languages other than English, Spanish, Portuguese or French were excluded.

BACKGROUND

Tuberculosis (TB) is one of the top 10 causes of death worldwide and a leading cause of death from an infectious disease. In 2018, 10 million people developed TB and 1.45 million people died from it globally, despite widespread availability of curative treatment.[1] Global treatment success was 85% for all new and relapse patients with TB in 2018. For HIV-associated TB, it was 75%. These proportions are lower than the End TB Strategy target of ≥90% treatment success.[2]

Heeding early recognition that *Mycobacterium tuberculosis* develops resistance rapidly in response to single-drug therapy, TB has been treated with combination regimens for more than 50 years.[3] Aside from weight-based dosing, the WHO and other TB guidelines authorities recommend a standardised approach for treatment of almost all patients with TB.[4–6] The current recommendation for drug-susceptible TB includes 2 months of isoniazid, rifampin, pyrazinamide, and ethambutol, followed by 4 months of isoniazid and rifampin.

Due to the long duration of TB treatment, it would be beneficial to understand early predictors of unsuccessful TB treatment outcomes to identify patients needing tailored treatment approaches, such as directly observed therapy (DOT) or extended treatment course. Research suggests that individual characteristics, such as HIV, age, undernutrition, diabetes, TB disease severity, extrapulmonary TB, history of TB, adherence, alcohol use and adverse drug reactions, are associated with unsuccessful TB treatment outcomes, but results vary by setting and patient population.[7–10]

Prediction models, defined as any combination or equation of two or more predictors to estimate an individualised probability of a specific endpoint within a defined period of time, are increasingly common in TB research.[11] The large number of recent prediction models for TB outcomes highlights the common desire to identify patients with TB at greatest risk of an unsuccessful treatment outcome. However, to date, there has not been a formal synthesis or quality assessment of existing prediction models for TB treatment outcomes, which is essential to determine whether they should be used to inform care and may help guide development of future models. Thus, we conducted a systematic review to identify, describe, compare and synthesise clinical prediction models designed to predict TB treatment outcomes among persons with pulmonary TB.

METHODS AND ANALYSIS

All steps of the systematic review were carried out according to guidelines set by Cochrane Prognosis Methods Group (PMG) and PROGnosis RESearch Strategy (PROGRESS).[12–14] Reporting adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, online supplemental file 1). This study was preregistered on Open Science Framework (OSF, https://osf.io/rz3wp) and the international prospective register of systematic reviews (PROSPERO: CRD42020155782).

Study eligibility criteria

The review question was defined according to the PICOTS (Population, Intervention, Comparator, Outcomes, Timing, Setting) framework (online supplemental file 2). In brief, the goal was to identify prognostic models developed to predict TB treatment outcomes among pulmonary TB cases. The main endpoint was unsuccessful TB treatment outcome, defined by the WHO as the combination of death, treatment failure, loss to follow-up and/or not evaluated, as compared with successful TB treatment outcome, defined as the combination of cure or treatment completion (table 1).[15] Loss to follow-up was sometimes referred to as default or treatment abandonment.

Inclusion criteria were: (1) prognostic model studies with or without external validation[16] ; (2) study population included adult, drug-susceptible, pulmonary, TB cases; (3) written in English, Spanish, Portuguese and French; (4) published between 1 January 1995 and 9 January 2020; (5) treatment outcome was one of the following: cure, treatment completion, death, treatment failure, loss to follow-up or not evaluated.

Exclusion criteria were: (1) predictive value of more than one variable was evaluated but not combined in a prediction model; (2) study population was only multidrug-resistant (MDR) TB cases, only extrapulmonary TB cases or only children (<18 years old); (3) outcome was evaluated during treatment such as: 2-month smear/culture conversion, acquired resistance, adverse events, quality of life; (4) long-term outcomes, such as relapse, recurrence or post-treatment mortality.

The decision to include only articles in English, Spanish, Portuguese and French was based on study team capabilities. The dates reflect modern TB treatment practice; first-line TB treatment regimens were not available until the early 1990s.[17 18] Articles that included a combination

Table 1 WHO definition of treatment outcomes for patients with TB

Outcome	Definition
Treatment completion	Completion of treatment without evidence of failure, but without documentation of a negative sputum smear or culture in the last month of treatment and/or on at least one previous occasion, either because tests were not done or because results are unavailable
Cure	Bacteriologic confirmation of a negative smear or culture at the end of TB treatment and on at least one previous occasion
Treatment success	Composite of cured and treatment completed
Treatment failure	Sputum smear or culture is positive at month 5 or later during treatment
Death	Patient with TB who dies for any reason before starting or during the course of treatment
Loss to follow-up	Patient with TB who did not start treatment or whose treatment was interrupted for 2 consecutive months or more
Not evaluated (transfer out)	Patient with TB for whom no treatment outcome was assigned, which includes cases who 'transferred out' to another treatment unit as well as cases for whom the treatment outcome is unknown to the reporting unit

TB, tuberculosis.

of drug-susceptible and drug-resistant cases, or a combination of children and adults were included.

Search strategy and selection criteria

The following electronic databases were searched on 9 January 2020: PubMed, Embase, Web of Science and the first 200 references from Google Scholar. This combination of databases achieved best overall recall for systematic reviews in a recent study.[19] Clinicaltrials.gov and retractiondatabase.org were also searched for unpublished research. Reference lists of retrieved articles were checked to identify eligible studies.

Search terms relating to the 'prediction model' component of the search were adapted from a PubMed search strategy that captured prediction model studies with sensitivity of 98%.[20] That component was combined with terms relating to TB treatment outcomes. The search strategy, developed in PubMed, was adapted for all other databases with assistance from a reference librarian (online supplemental file 3).

Article selection was conducted in three stages. The first stage was automatic deduplication and title screening, carried out using *revtools* in RStudio (V.1.2).[21] Remaining articles were imported into Covidence, a web-based software platform that streamlines systematic reviews, where abstracts (Stage 2) and full text (Stage 3) were manually screened.[22] Stages 2 and 3 were carried out by two independent reviewers (LP and FR). Discordance was discussed between reviewers, and if consensus was not reached, a third party arbitrated (one of TS, VCR, PR, DL). In stage 3, reasons for exclusion were documented according to PRISMA.

Data analysis

Data from selected studies were recorded using a database designed in REDCap (Vanderbilt University).[23 24] Data extraction was informed by the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) and the Prediction Model Risk of Bias Assessment Tool (PROBAST).[16 25 26] CHARMS checklist and PROBAST are shown in online supplemental files 4 and 5, respectively.

Quality assessment and applicability of included studies was assessed using PROBAST by dual independent review.[16 26] PROBAST was specifically designed to assess risk of bias of prediction model studies, which included identifying deficiencies in study design, conduct or analysis that led to inaccurate estimates of predictive performance. PROBAST has four domains: participants, predictors, outcome and analysis with 20 total signalling questions. Each question was answered on the scale: yes, probably yes, no, probably no, no information. Domains were scored as low, high and unclear risk of bias. PROBAST also guides assessment of applicability of participants, predictors, and outcomes from each included study to the review question.

Results were summarised narratively and in tables and figures. Meta-analysis was not possible due to lack of external validation and use of disparate predictors, outcome definitions and modelling methods. For studies that presented multiple models with the same set of predictors and outcomes, but different methods, the best-performing method was included in data synthesis. For studies presenting multiple models with different sets of predictors (ie, baseline data vs longitudinal data), the model developed using only baseline data was included. If studies developed multiple models for different outcomes or with different populations, all models were included. To further evaluate the impact of study population heterogeneities on prediction model performance, we additionally examined results after stratifying studies by inclusion/exclusion of MDR and younger age groups.

Patient and public involvement

Neither patients nor the public were involved in the design, conduct, or reporting of the research, as it was not feasible or appropriate for this systematic review. The study protocol is publicly available at https://osf.io/rz3wp.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

RESULTS

Study selection

The search identified 14 739 unique studies. After excluding irrelevant titles, 6426 abstracts were screened, 536 articles underwent full-text review, and 33 model development studies presenting 37 prediction models were included (figure 1).

Study characteristics

Of the 33 studies, most were retrospective cohorts (n=25, 76%), three (9%) were prospective cohort studies, two (6%) were case–control studies and three (9%) were nested case–control studies. Data from nearly half of studies (n=16, 48%) were collected from surveillance systems; 11 (33%) studies used a data collection form developed specifically for their study and 6 (18%) studies extracted data from medical records. Median sample size was 803 (IQR: 291–4167). Full details on included studies are in table 2.

Thirteen (41%) studies took place in Asia, eight (25%) in Africa, six (19%) in Europe, four (12%) in North America and one (3%) included sites in Europe and Argentina. Fewer than half (n=14, 45%) took place in high-burden TB settings.[1] One study did not report study location (tables 2 and 3).

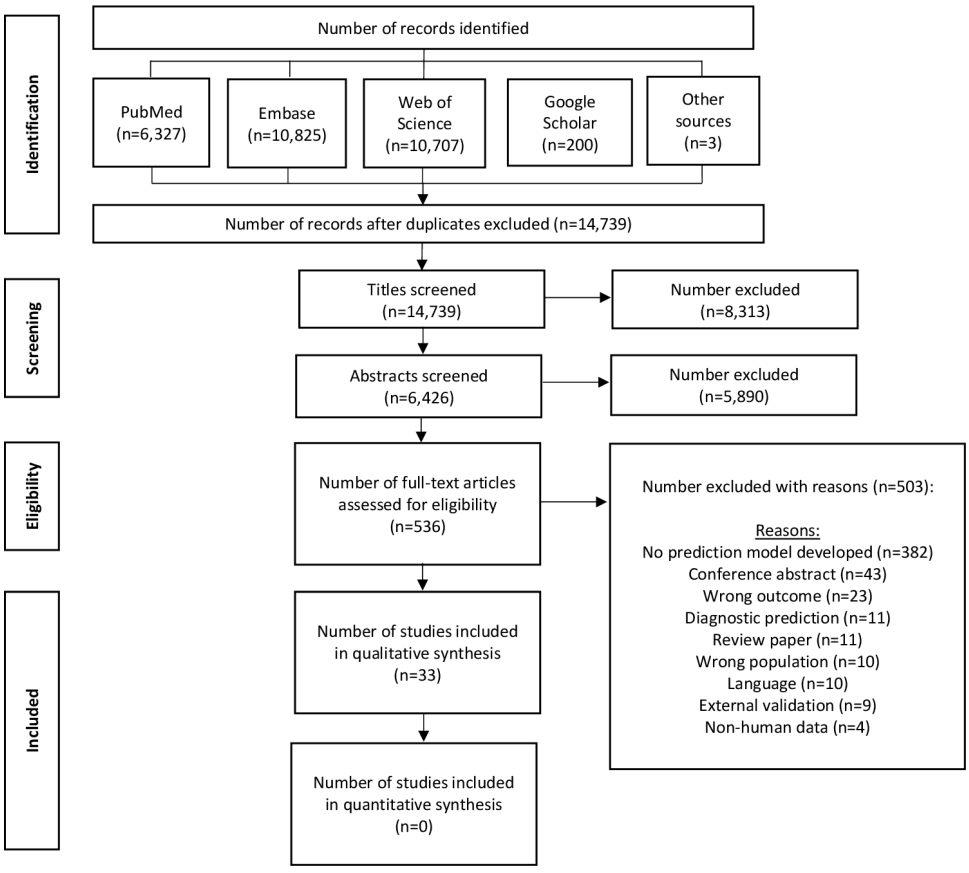

Figure 1 PRISMA flow chart of inclusion process. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Reporting of population characteristics varied by study (table 4). Among 18 studies that reported a measure of central tendency (mean or median) for age, the median of those measures was 41 years (IQR: 37–49). Of 17 studies that reported the minimum age of participants, seven (41%) had a minimum age of 15, one (6%) had a minimum age of 16, one (6%) had a minimum age of 17 and the remainder had minimum age of 18. Eighteen studies reported including persons living with HIV (PLWH); 5 of these included only patients with TB/HIV. Thirteen studies reported including persons with diabetes; one of which included only TB/DM. Eight studies reported including some participants with MDR, though prevalence of MDR was low in all studies. Ten studies included only hospitalised patients, and in 14 studies, all participants were on directly observed therapy (DOT).

Model characteristics

Model outcomes included death (n=16, 43%), treatment failure (n=6, 16%), default (n=6, 16%) or a composite outcome (n=9, 25%, tables 2 and 5). The complete outcome definition for all included studies is in online supplemental file 6.

Most models were developed using clinical/epidemiologic predictors (n=34, 92%), two (5%) used multiple biomarkers and one (3%) used adherence data. The most common candidate predictors were age, sex, extrapulmonary TB, smear result, body mass index (BMI), X-ray findings and previous TB. The most common predictors retained in the final models were age, sex, extrapulmonary TB, BMI, chest X-ray results, previous TB and HIV (figure 2).

Only three models (8%) used survival analysis; most models used logistic regression (n=29, 78%) and five (14%) used a machine-learning approach. More than half of studies (n=19, 51%) considered variables for inclusion in the multivariable model based on unadjusted associations with the outcome. Model building methods varied widely between models (table 5).

Only 19 (51%) models were internally validated, including 10 (53%) split-sample validation, 5 (26%) bootstrap resampling and 4 (21%) cross-validation. Six (16%) models were externally validated. Many models (n=30, 81%) reported discrimination with c-statistic (concordance statistic) or area under the receiver operating characteristic (AUROC), which are equivalent and quantify the ability of the model to distinguish between patients who do and do not develop an outcome. Only 17 (46%) reported calibration, the agreement between observed and predicted outcomes. Most studies assessed calibration with Hosmer-Lemeshow tests (n=13, 77%); only two studies provided a calibration plot, the preferred reporting method for prediction model studies,[16 27 28] and one reported the calibration slope (table 2). Models were

Table 2 Study characteristics

First author, year	Population	Study years	Study design	Location	Validation	No. with outcome/sample size (%)	Predictors in final model	Performance measures	Model presentation	Risk of bias (population, predictor, outcome, analysis)
Death										
Abdelbary *et al*[β]/2017	TB cases	2006–2013	Retrospective cohort	Mexico	Internal (split-sample)	*Development:* 261/4216 (6%) *Validation:* 260/4215 (6%)	Age (<41, 41–65, ≥65), sex, MDR, HIV, malnutrition, alcoholism, diabetes, pulmonary TB	c-statistic=0.70 Sensitivity=60% Specificity=71%	Risk score	Low, high,low, high
Abdelbary *et al*[β]/2017 (TB/DM)	TB/DM cases	2006–2013	Retrospective cohort	Mexico	None	88/2121 (4%)	Sex, malnutrition, BCG vaccinated, AFB smear (positive vs negative)	c-statistic=0.68	Risk score	Unclear, high, low, high
Aljohaney[69]/2018	Hospitalised patients with TB	December 2011 – December 2016	Retrospective cohort	Saudi Arabia	None	41/291 (14%)	*Clinical model*: age, congestive heart failure *Clinical + lab model:** age >65, congestive heart failure, bilateral disease on chest X-ray	*Clinical model*: Accuracy=86% *Clinical and lab model:** Accuracy=90%	ORs	Unclear, unclear, unclear, high
Bastos *et al*[70]/2016	Inpatient and outpatient TB cases on DOT	2007–2013	Retrospective cohort	Portugal	External (setting)	*Development:* 121/681 (18%) *Validation:* 24/103 (23%)	Hypoxemic respiratory failure, age (≥50 vs <50), bilateral involvement, comorbidities (at least one of HIV, diabetes, liver failure/cirrhosis, congestive heart failure, chronic respiratory disease), haemoglobin (<12 vs ≥12)	AUROC=0.84 (95% CI: 0.76 to 0.93) Sensitivity=41.8% Specificity=92.1%	Risk score	Low, unclear, low, high
Gupta-Wright *et al*[44]/2019	Hospitalised patients with TB/HIV	October 2015–September 2017	Retrospective cohort	Malawi and South Africa	External (setting)	*Development:* 94/315 (30%) *Validation:* 147/644 (23%)	Sex, age 55+, currently taking ART, ability to walk unaided, severe anaemia, positive TB-LAM	c-statistic=0.68 (95% CI: 0.61 to 0.74) HL test: p=0.13 Calibration plot	Risk score	Low, low, low, high
Horita *et al*[71]/2013	Hospitalised patients with TB	January 2008–July 2011	Retrospective cohort	Japan	External (setting)	*Development:* 36/179 (20%) *Validation:* 48/244 (20%)	Age, oxygen requirement, albumin, activities of daily living	AUROC=0.893 Sensitivity=0.92 Specificity=0.73	Risk score	Low, low, low, high
Koegelenberg *et al*[40]/2015	Hospitalised patients with TB	January 2012–May 2013[66]	Retrospective cohort	South Africa	None	38/83 (46%)	Septic shock, HIV with CD4 <200, creatinine >140 (male) or >120 (female), P:F O2 ratio <200, chest radiograph showing miliary pattern/parenchymal infiltrates, absence of TB treatment at admission	Mean score in survivors: 2.27 (SD=1.47) Mean score in non-survivors: 3.58 (SD=1.08)	Risk score	Low, Low, Low, High
Nguyen and Graviss[53](general pop)/2018	TB cases	January 2010–December 2016	Retrospective cohort	Texas	Internal (split-sample)	*Development:* 253/3378 (7%) *Validation:* 270/3377 (8%)	Age group (15–44, 44–64, >64), US born, homeless, resident of long-term care facility, chronic kidney failure, meningeal TB, miliary TB, HIV positive, HIV unknown	AUROC=0.80 (95% CI: 0.77 to 0.82) HL test: χ^2=6.3, p=0.613	Risk score	Low, unclear, unclear, high
Nguyen and Graviss[37](TB/DM)/2019	Patients with TB/DM	January 2010–December 2016	Retrospective cohort	Texas	Internal (bootstrap)	112/1227 (9%)	Age ≥65, US-born, homeless, injection drug use, chronic kidney failure, TB meningitis, Miliary TB, AFB positive smear, HIV positive	AUROC=0.82 (95% CI: 0.78 to 0.87) HL test: χ^2=4.54, p=0.81 Brier score=0.07	Risk score	Unclear, unclear, unclear, high
Nguyen *et al*[52] (TB/HIV)/2018	Patients with TB/HIV	January 2010–December 2016	Retrospective cohort	Texas	Internal (bootstrap)	57/450 (13%)	Age ≥45, resident of long-term care facility, meningeal TB, abnormal chest x-ray, diagnosis confirmed by positive culture of nucleic acid amplification, culture not converted or unknown	AUROC=0.79 (95% CI: 0.70 to 0.87) HL test: χ^2=4.25, p=0.51 Brier score: 0.09	Risk score	Low, high, unclear, high
Pefura-Yone *et al*[54]/2017	Patients with TB	January 2012–December 2013[66]	Retrospective cohort	Cameroon	Internal (bootstrap)	213/2250 (9%)	Age, adjusted BMI, clinical form (smear-positive pulmonary TB, Psmear-negative pulmonary TB, extrapulmonary TB), HIV	C-statistic: 0.808 HL test: χ^2=6.44, p=0.60 Sensitivity=80.7% Specificity=68.2% Calibration plot	Model coefficients	Low, low, low, high

Continued

Table 2 Continued

First author, year	Population	Study years	Study design	Location	Validation	No. with outcome/sample size (%)	Predictors in final model	Performance measures	Model presentation	Risk of bias (population, predictor, outcome, analysis)
Podlekareva *et al*[46]/2013	Patients with TB/HIV	January 2004–December 2006	Retrospective cohort	52 cities in Europe and Argentina	None	995†	Drug susceptibility testing performed, treatment with rifamycin+isoniazid+pyrazinamide, and combination ART at/near TB diagnosis	Crude hazard ratio=0.62 (95% CI: 0.64 to 0.84)	Risk score	Low, unclear, low, high
Valade *et al*[42]/2012	Hospitalised patients with TB	March 2000–July 2009	Retrospective cohort	France	Internal (bootstrap)	20/53 (38%)	Miliary TB, catecholamine infusion, mechanical ventilation on admission	AUROC=0.92 (95% CI: 0.85 to 0.98) Brier score=0.13 Optimism=0.03 Accuracy=85% Sensitivity - 75% Specificity=91%	Risk score	Unclear, low, low, high
Wang *et al*[43]/2019	HIV-negative, culture-confirmed, pulmonary TB cases	January 2014–December 2016	Prospective cohort	China	External (setting)	*Development:* 36/287 (13%) *Validation:* 15/104 (14%)	Age, cavitary lesion, pleural effusion, drug resistance, disseminated, albumin, c-reactive protein, white blood cell count, IL-6, migration inhibitory factor	AUROC=0.85 ± 0.028	ORs	Low, low, low, high
Wejse *et al*[36]/2008	Patients with Pulmonary TB on DOT	1996–2001	Retrospective cohort	Guinea Bissau	None	100/698 (14%)	Cough, haemoptysis, dyspnoea, chest pain, night sweating, anaemia conjunctivae, tachycardia, positive funding at lung auscultation, temperature >37, BMI <18, BMI <16, mid-upper arm circumference (MUAC) <220, MUAC <200	AUROC=0.65 (95% CI: 0.6 to 0.7) Sensitivity=0.45 Specificity=0.75	Risk score	Low, high, low, high
Zhang *et al*[45]/2019	Patients with TB/HIV at end stage of AIDS	August 2009–January 2018	Retrospective cohort	China	Internal (split-sample)	*Development:* 157/807 (19%) *Validation:* 40/200 (20%)	Anaemia, TB meningitis, severe pneumonia, hypoalbuminaemia, unexplained infection or space-occupying lesions, malignancy	AUROC=0.867 (95% CI: 0.832 to 0.902) Sensitivity=79.6% Specificity=82.9%	Risk score	Low, low, low, high
Treatment failure										
Abdelbary *et al*[9]/2017	TB cases	2006–2013	Retrospective cohort	Mexico	Internal (split-sample)	*Development:* 2109† *Validation:* 6322†	Education (no or low vs higher than primary school), MDR, AFB smear (>+2,+1, negative)	c-statistic=0.65 Sensitivity=52% Specificity=66%	Risk score	Low, high, low, high
Kalhori *et al*[49] (logistic)/2010	TB cases at DOTS registration	2005	Retrospective cohort	Iran	Internal (split-sample)	*Development:* 828/4836 (17%) *Validation:* 2418†	Gender, age, weight nationality, prison, case type	AUROC=0.70 Accuracy=81.64% HL test: χ^2=11.935, df=8, p=0.154	Model coefficients	Unclear, unclear, unclear, high
Keane *et al*[30]/1997	Patients with smear-positive TB on standard first-line regimen with DOT	1990–1995	Non-nested case–control	Vietnam	None	130/803 (16%)	*3 month model*: extensive lesions, mediastinal shift, average smear score third month, weight, progressive X-ray, any previous treatment *Baseline model*: mediastinal shift, average smear score, extensive lesions, any previous treatment, cavities, weight	*3 month*: Sensitivity=80% Specificity=80% *Baseline*: Sensitivity=70% Specificity=80%	Model coefficients	High, unclear, unclear, high
Luies *et al*[33]/2017	Smear-positive pulmonary TB cases on DOT	May 1999–July 2002	Nested case–control	South Africa	Internal (cross-validation)	10/31 (32%)	3,5,-Dihydroxybenzoic acid, (3-(4-Hydroxy-3-methoxyphenyl) propionic acid	AUROC=0.89 (95% CI: 0.7 to 1.00)	Model coefficients	High, unclear, unclear, high
Mburu *et al*[72]/2018	Patients with smear-positive TB	February 2014–August 2015	Prospective cohort	Kenya	Internal (cross-validation)	13/321 (4%)	HbA1c, regimen (retreatment), age, weight, random blood glucose, BMI, blood urea nitrogen, HIV-positive result, ever smoker, creatinine	AUROC=0.56 ± 0.07	Relative score	Low, low, low, high
Default										
Thompson *et al*[73]/2017	Adults who were HIV-uninfected with newly diagnosed pulmonary TB	April 2010–April 2013	Retrospective cohort	South Africa	Internal (cross-validation) and external (setting)	6/99 (6%)	18 splice junctions and 13 genes	AUROC (internal)=0.87 AUROC (external)=0.63	Heatmap of differentially expressed genes	Low, low, low, high

Continued

Table 2 Continued

First author, year	Population	Study years	Study design	Location	Validation	No. with outcome/sample size (%)	Predictors in final model	Performance measures	Model presentation	Risk of bias (population, predictor, outcome, analysis)
Abdelbary *et al*[9]/2017 (TB/DM)	TB cases	2006–2013	Retrospective cohort	Mexico	None	93/2121 (4%)	Age (<40 vs ≥40), sex, HIV	c-statistic=0.62	Risk score	Unclear, high, unclear, high
Belilovsky *et al*[35]/2010	Hospitalised patients with TB	1993–2002	Retrospective cohort	Russia	External (geographical)	*Development:* 1326/3904 (34%) *Validation:* 4662/12803 (36%)	Sex, unemployment, retreatment case, alcohol abuse (yes, no, no data), severe TB form, residence (urban vs rural), age (25–50 vs other), pulmonary TB (vs extrapulmonary), prison history	Belgrood: AUROC=0.75 Orel: AUROC=0.75 Pskov: AUROC=0.78 Yaroslavi: AUROC=0.75 Calibration table	Model coefficients	Unclear, high, high, high
Chang *et al*[31]/2004	All patients with TB	January 1999–March 1999	Nested case–control	China	None	102/408 (25%)	*Baseline*:* ever smoker (current, former, never), retreatment (history of default, no history of default, not) *Longitudinal*: smoking status (current, former, never), retreatment (with history of default, without history of default, never), unsatisfactory adherence in first 2 months (good, poor, fair, unknown), subsequent hospitalisation, treatment side effects in last month of treatment	Baseline:* AUROC=0.70 (95% CI: 0.63 to 0.76) HL test: χ^2=1.448, df=5, p=0.919 *Longitudinal*: AUROC=0.85 (95% CI: 0.80 to 0.90) HL test: χ^2=5.887, df=6, p=0.436	ORs	High, high, low, high
Chee *et al*[32]/2000	TB cases	1996	Nested case–control	Singapore	None	38/71 (54%)	Chinese race, extent of family support, treatment duration	Accuracy=74.6%	Model coefficients	High, unclear, high, high
Cherkaoui *et al*[29]/2014	Patients with TB with definite or probable pulmonary or extrapulmonary TB	June 2010–October 2011	Non-nested case–control	Morocco	None	91/277 (33%)	Age <50, work interfering with ability to take TB treatment, retreatment regimen, daily DOT, moderate or severe side effects, told friends about TB, current smoker, never smoker, symptom resolution in <2 months, knowledge of TB treatment duration	AUROC=0.85 (95% CI: 0.80 to 0.90) Sensitivity=82.4% Specificity=87.6% HL test: χ^2=0.77, p value=1.00	Survey tool	High, high, high, high
Rodrigo *et al*[55]/2012	New TB cases	January 2006–December 2009	Prospective cohort	Spain	Internal (split-sample)	*Development:* 92/1490 (6%) *Validation:* 103/1589 (6%)	Immigrant, living alone, living in an institution, previous TB treatment, linguistic barriers (poor understanding), intravenous drug use, unknown intravenous drug use	AUROC=0.67 (95% CI: 0.65 to 0.70) Sensitivity=65.05% Specificity=67.36%	Risk score	Low, low, low, high
Unfavourable outcome										
Kalhori and Zeng[50](predicting)/2009†	Patients with TB at DOT registration	2005	Retrospective cohort	Iran	Internal (split-sample)	*Development:* 6920† *Validation:* 2966†	Age, gender, nationality, prison, area, weight	Classification rate=89.8% R2=0.45	Model coefficients	Unclear, unclear, unclear, high

Continued

Table 2 Continued

First author, year	Population	Study years	Study design	Location	Validation	No. with outcome/sample size (%)	Predictors in final model	Performance measures	Model presentation	Risk of bias (population, predictor, outcome, analysis)
Sauer *et al*[57]/ 2018†	TB cases	Data available through March 2018	Retrospective cohort	Azerbaijan, Belarus, Georgia, Moldova, Romania	Internal (split-sample)	*Development:* 103/411 (25%) *Validation:* 44/176 (25%)	*FS:** Drug sensitivity, employment status, smear microscopy, dissemination Backwards elimination (BE): Drug sensitivity, employment status, smear microscopy, dissemination Stepwise selection (SS): Drug sensitivity, employment status, smear microscopy, dissemination Lasso: Country, employment, extrapulmonary, cavity size, decrease in lung capacity, smear microscopy, drug sensitivity, chest imaging Random forest (RF): Top five by mean decrease accuracy: lung cavity size, type of resistance, employment status, country, total cavities Top five by mean decrease Gini index: Age of onset, drug regimen, lung cavity size, number of daily contacts, culture	*FS:** AUROC=0.74 (95% CI: 0.66 to 0.82) Sensitivity=0.36 Specificity=0.89 Misclassification=0.24 BE: AUROC=0.73 (95% CI: 0.65 to 0.81) Sensitivity=0.3 Specificity=0.88 Misclassification=0.27 SS: AUROC=0.73 (95% CI: 0.65 to 0.81) Sensitivity=0.30 Specificity=0.88 Misclassification=0.27 Lasso: AUROC=0.72 (95% CI: 0.64 to 0.80) Sensitivity=0.21 Specificity=0.96 Misclassification=0.23 RF: AUROC=0.73 (95% CI: 0.65 to 0.81) Sensitivity=0.30 Specificity=0.88 Misclassification=0.27 SVM linear: AUROC=0.69 (95% CI: 0.60 to 0.77) Sensitivity=0.21 Specificity=0.94 Misclassification=0.24 SVM polynomial: AUROC=0.69 (95% CI: 0.60 to 0.77) Sensitivity=0 Specificity=1 Misclassification=0.25	List	Unclear, unclear, unclear, high
Baussano *et al*[47]/2008‡	Pulmonary TB cases	2001–2005	Retrospective cohort	Italy	Internal (bootstrap)	576/1242 (46%)	Residency (residential vs homeless), sex, geographic origin (non-EU vs EU), case definition (other than definite vs definite), treatment setting (inpatient and unknown vs outpatient), age (continuous)	AUROC=0.75 Calibration slope=0.98 R^2=0.24	Nomogram	Low, unclear, low, high
Costa-Veiga *et al*[48]/2017‡	Pulmonary TB cases	2000–2012	Retrospective cohort	Portugal	External (temporal)	*Development:* 1152/10766 (11%) *Validation:* 4714†	HIV, previous treatment, age class (25–44, 15–24, 45–64,>64), intravenous drug use, pathologies (other disease comorbidity)	AUROC=75.9% (95% CI: 74.1 to 77.7) Sensitivity=71% Specificity=73%	Nomogram	Low, low, low, high

Continued

Table 2 Continued

First author, year	Population	Study years	Study design	Location	Validation	No. with outcome/sample size (%)	Predictors in final model	Performance measures	Model presentation	Risk of bias (population, predictor, outcome, analysis)
Killian *et al*[34]/2019‡	Patients with TB (99 DOTS programme)	February 2017–September 2018	Retrospective cohort	India	None	433/4167 (10%)	*LEAP*:* LEAP with two input layers, (1) LSTM with 64 hidden units and a dense layer with 48 units for the dense layer and four units for the penultimate layer w-misses: missed doses in last week t-misses: total missed doses in 35 days units and a dense layer with 48 units for the dense layer and four units for the penultimate layer Random forest: 150 trees and no max depth based on digital adherence technology from first 35 day	*LEAP** AUROC=0.743 *lw-misses*: AUROC=0.607 *t-misses*: AUROC=0.630 *Random forest*: AUROC=0.722	None	High, high, unclear, high
Madan *et al*[51]/2018‡	Patients with TB/HIV/HIV on DOT with first-line TB treatment	2015	Retrospective cohort	India	None	78/448 (17%)	Sputum smear grade, previous TB, disease classification, HIV status, ART status, CD4 cell count, sex and age group (with interaction terms between age group and sex; sputum smear status and type of TB; HIV status at TB diagnosis and CD4 cell category).	AUROC=0.783 HL test p value=0.149	Model coefficients	Low, low, low, high
Mburu *et al*[72]/ 2018‡	Patients with Smear-positive TB	February 2014–August 2015	Prospective cohort	Kenya	Internal (cross-validation)	32/340 (9%)	HbA1c, treatment regimen (retreatment), creatinine, BMI, blood urea nitrogen, weight, age, random blood glucose, HIV positive result, male gender	AUROC=0.65 ± 0.06	Relative score	Low, low, low, high
Other outcome										
Kalhori and Zeng[74](fuzzy)/2009§	Patients with TB at DOTS registration	2005	Retrospective cohort	Iran	Internal (split-sample)	*Development:* 7254† *Validation:* 2418†	Case type, treatment category, risky sex, prison, sex, recent TB infection, diabetes, low body weight, TB type, length, previous imprisonment, age, area, HIV	Mean absolute percentage error=1.24	Learnt parameters	Unclear, unclear, high, high
Hussain and Junejo[56]/2019¶	Patients with pulmonary and extrapulmonary TB (TB Reach)	2011–2014	Retrospective cohort	Unknown	Internal (split-sample)	*Development:* 3371† *Validation:* 842†	Random forest*, artificial neural networks and support vector machine	*Random forest*:* Accuracy=76.32%	None	Unclear, unclear, unclear, high

*Indicates best-performing/most relevant model, which is included throughout the manuscript (see Methods section for details). Performance measures are reported for highest level of validation performed (ranked from strongest to weakest: external validation, internal validation, no validation). If internal and external validation were performed, both are reported.
†Outcome number unknown.
‡Outcome is composite of death, treatment failure, loss to follow-up and not evaluated.
§Outcome is a value from 1 to 5 (1=patient completed the treatment course in frame of DOTS, 2=cured, 3=quit treatment, 4=failed treatment and 5=death).
¶Outcome is treatment completion.
**Outcome is composite of death and treatment failure (losses to follow-up and not evaluated (unknown) outcomes were excluded).
AFB, acid fast bacilli; AUROC, area under receiver operating characteristic; BCG, Bacillus Calmette–Guérin; BMI, body mass index; c-statistic, concordance statistic; DM, diabetes mellitus; DOTS, directly observed therapy; FS, forward selection; HbA1c, haemoglobin A1c; HL, Hosmer-Lemeshow; LEAP, Lstm rEal-time Adherence Predictor; MDR, multidrug resistant; TB, tuberculosis.

Table 3 Characteristics of patient populations in the 33 included studies with prediction models for TB treatment outcomes

Characteristics	Studies reporting characteristic, n (%)	Categories	N (%) or median (IQR)
Sample size	33 (11)	–	803 (291–4167)
Study duration, years	32 (97)	–	4 (2–7)
Study design	33 (100)	Prospective cohort	3 (9)
		Retrospective cohort	25 (76)
		Nested case–control	3 (9)
		Non-nested case–control	2 (6)
Data source	33 (100)	Medical record	6 (18)
		National registry or surveillance system	13 (39)
		Local registry or surveillance system	1 (3)
		Regional registry or surveillance system	2 (6)
		Data collect form for study purposes	11 (33)
Study region	32 (97)	Africa	8 (25)
		Asia	13 (41)
		Europe	6 (19)
		North America	4 (12)
		South America	0 (0)
		Global	1 (3)
High burden TB setting*	31 (94)	All	143 (42)
		Some	1 (3)
		None	17 (55)
Missing data	18 (54)	Complete case analysis	9 (50)
		Missing indicator method	4 (22)
		Heckman's method	1 (6)
		Simple imputation	2 (12)
		Sensitivity analysis with imputation	1 (6)
		Other	1 (5)
Number of models developed	33 (100)	1	25 (76)
		2	4 (12)
		3	1 (3)
		4	2 (6)
		7	1 (3)
Reasons for multiple models developed	8 (24)	Different outcomes	1 (12)
		Different predictors considered	4 (50)
		Different methods	2 (25)
		Different outcomes	1 (12)
		Different populations and outcomes	1 (12)

*Determined based on study location and WHO list of 30 countries with high-burden TB in the 2019 Global Tuberculosis Report (1).
TB, tuberculosis.

presented a variety of ways, the most common of which was a weighted risk score (n=16, 43%); details on model presentation are in online supplemental file 7.

Quality assessment

Grading of PROBAST signalling questions is summarised in figure 3, and the summary risk of bias for the participants, predictors, outcome and analysis domains and assessment of applicability are shown in figure 4. More than half of the studies were at low risk of bias for the population and outcomes domains, but all studies were at high risk of bias in the analysis domain.

Table 4 Study population characteristics of 33 included studies

Characteristics	Included? Yes	No	Unknown	Median (IQR)*, n
Age†	–	–	15	41 (37–49), n=18
HIV	18	7	8	23% (10–100), n=17
Diabetes	13	1	19	12% (5–21), n=11
MDR	8	7	18	1% (1–3), n=8
Other drug resistance	12	1	20	6% (4–12), n=10
Extrapulmonary TB‡	22	4	7	11%(4–17), n=16
Previous TB	20	1	12	19% (9–30), n=17
DOT	14	0	19	100% (100–100), n=14
Hospitalised patients	13	1	19	100% (100–100), n=10

*Other than age (which is reported in years), this is the percentage of the population that has the characteristic among studies that include patients with the characteristic. For example, among the 18 studies that include persons with HIV, 17 report how many people had HIV and among those, the median percentage of the population with HIV is 23%.
†Based on the measure of central tendency reported in the study (mean: n=11; median: n=7).
‡Forms of extrapulmonary TB differ by study but included some of the following: miliary, meningeal, pleural, peritoneal, disseminated, blood/bone, abdominal.
DOT, directly observed therapy; MDR, multidrug resistance; TB, tuberculosis.

Common sources of population bias included use of non-nested case–control design,[29 30] nested case–control design without proper estimation of baseline risk,[31 32] or inappropriate inclusion/exclusion criteria.[33 34] Sources of predictor bias included lack of standardised assessment of key predictors (ie, HIV, diabetes, chest X-ray scoring)[9 29 31 34–36] or timing of data collection/availability that would limit the intended use of the model.[9 29 37] Within the outcomes domain, sources of bias included subjective[35] or non-standard[32 38] outcome measures and inconsistent outcome ascertainment.[29]

Bias in the analysis domain was widespread. More than half of the models included were likely overfit due to low events per variable ratios (table 5). Only six studies handled continuous and categorical variables appropriately (ie, did not dichotomise continuous variables, considered non-linearity of continuous variables).[31 39–43] Most studies used complete case analysis or did not mention missing data; no study used multiple imputation in their main analysis. One study with low amounts of missing data (<5%) conducted sensitivity analysis with multiple imputation.[44] A different study excluded only two people out of a total sample size of 1007 with missing data, which would have little impact on model performance.[45] Fewer than half (n=14) of studies avoided univariable predictor selection, and only three studies used survival analysis, appropriately accounting for censoring.[36 45 46] Performance measures were appropriately reported (ie, calibration assessed with plot and discrimination assessed with c-statistic/AUROC) in three studies.[41 44 47] Only two studies estimated optimism (degree to which data are overfit) or accounted for potential overfitting with penalisation of model parameters.[35 41] Ten studies appropriately presented their model with model coefficients or nomograms, which prevents bias from rounding or transforming model coefficients to generate a risk score.[30 33 35 37 38 45 47–55]

About half of the models (n=19, 51%) were applicable to the review question in all domains. However, unclear reporting of target population or predictor and outcome definitions limited assessment of applicability for several studies.[38 49 50 56 57] Additionally, studies that included only hospitalised patients with specific laboratory parameters may not be routinely available in the clinical setting.[39 40 42] Results from analyses stratified by inclusion of patients with MDR and minimum age <18 are presented in online supplemental file 8.

DISCUSSION

In this comprehensive, systematic review of prediction models for pulmonary TB treatment outcomes, we identified 33 model development studies presenting 37 prediction models. Although diagnostic prediction models for prevalent TB were previously systematically reviewed, this is the first review of TB treatment outcomes.[58] The included prediction models were developed for predicting death, treatment failure, default or a composite unfavourable outcome during TB treatment. Most models reported good performance (c-statistic/AUROC >0.7), but all were evaluated to have high risk of bias due to poor reporting, exclusion of missing data, weak methodologic approaches, lack of calibration assessment and limited validation. Population heterogeneities, such as differences in inclusion/exclusion of individuals with MDR and younger ages, and varying predictor and outcome definitions limited comparisons between models.

Table 5 Methods reported for the 37 models of the 33 included studies with prediction models for TB treatment outcomes

Characteristics	Studies reporting characteristic, n (%)	Categories	N (%) or median (IQR)
Type of outcome	37 (100)	Single	29 (78)
		Composite	8 (22)
Outcome	37 (100)	Death	16 (43)
		Treatment failure	6 (16)
		Default, loss to follow-up or treatment interruption	6 (16)
		Unfavourable outcome	6 (16)
		Treatment success	2 (6)
		Other*	1 (3)
Number—prevalence of outcome†	32 (87)	–	94 (38–171) 15% (9-–26)
Events per candidate variable‡	30 (81)	–	6 (3-–11)
Events per variable (in final model)	29 (78)	–	14 (9–26)
Predictor types	37 (100)	Clinical/epidemiologic	34 (92)
		Adherence	1 (3)
		Biomarker	2 (5)
Analysis	37 (100)	Logistic regression	29 (78)
		Survival analysis	3 (8)
		Machine learning	5 (14)
Method for considering predictors in multivariable models	36 (97)	All candidate predictors	12 (32)
		Based on unadjusted association with outcome	19 (51)
		Based on clinical relevance	1 (3)
		Other§	4 (14)
Selection of predictors during modelling	31 (84)	Full model approach	2 (6)
		Forward selection	7 (23)
		Backwards elimination	5 (16)
		Stepwise selection	8 (26)
		Random Forest	1 (3)
		Hosmer-Lemeshow model building criteria	4 (13)
		Bayesian model averaging	3 (10)
		Pairwise selection	1 (3)
P value for consideration in model	17 (46)	0.01	2 (12)
		0.05	3 (18)
		0.11	1 (6)
		0.2	6 (35)
		0.25	5 (29)
P value for retention in MV model	20 (54)	0.05	9 (45)
		0.1	9 (45)
		0.15	1 (5)
		0.2	1 (5)
Internal validation	19 (51)	Split-sample	10 (53)
		Bootstrap	5 (26)

Continued

Table 5 Continued

Characteristics	Studies reporting characteristic, n (%)	Categories	N (%) or median (IQR)
		Cross-validation	4 (21)
External validation	6 (16)	Temporal	1 (17)
		Geographic	1 (4)
		Setting	4 (67)
Calibration	17 (46)	Calibration plot¶	2 (12)
		Calibration slope¶	1 (6)
		Hosmer-Lemeshow goodness of fit p value¶	13 (77) 0.51 (0.20–0.79)
		Calibration table¶	2 (12)
		Mean absolute error¶	1 (6)
Discrimination	30 (81)	C-statistic (AUROC)¶	30 (100) 0.75 (0.68–0.84)
		Log rank test¶	2 (5)
Classification	18 (49)	Sensitivity**	14 (78) 70(54, 78)
		Specificity**	13 (72) 75 (71–88)
		Accuracy	2 (11)
		Other††	2 (11)
Model presentation	34 (92)	Risk score	16 (43)
		Model coefficient	8 (22)
		Nomogram	2 (6)
		ORs/relative scores	4 (12)
		Survey tool	1 (3)

*Outcome is a value from 1 to 5 (1=patient completed the treatment course in frame of DOTS, 2=cured, 3=quit treatment, 4=failed treatment and 5=death).
†Prevalence of outcome in the population used to develop the prediction model (ie, derivation/development subset if split-sample technique was used or full sample if the model was not validated or if bootstrap/cross-validation was used).
‡Only five studies report the exact number of predictors considered. Otherwise, the number of candidate predictors was estimated from the provided tables or lists of candidate predictors in the source paper.
§Other methods of determining which variables to consider for prediction model include: principal components analysis (n=1), screening for multicollinearity via correlation coefficient (n=1), one study used a combination of a priori and selection via univariable association, and the other used machine-learning preprocessing (n=1).
¶Sums to more than 100%, because some studies report multiple measures of calibration or discrimination.
**Based on the following cut-off methods: Youden (n=4) concordance probability (n=1), estimated at nearest 0,1 for studies that present a range of sensitivity and specificity in a table or figure (n=4), or unknown (n=5).
††Other includes one study that reports false positive rate and one study that includes a graph of sensitivity versus specificity.
AUROC, area under receiver operating characteristic; c-statistic, concordance statistic; TB, tuberculosis.

More than half of the models included in the review were developed in low-burden TB settings, and none were developed specifically in South America. Prediction of TB treatment outcome is especially important in high-burden TB settings, where resources may be limited, and risk assessment can guide resource allocation toward patients who need the most involved care.

Common risk factors included in the models were consistent with well-established risk factors for poor TB treatment outcomes, including age, sex, HIV, extrapulmonary TB, baseline smear results and previous TB treatment. Among studies that included PLWH, only three considered factors related to management/severity of HIV, such as receipt of antiretroviral therapy, CD4 cell count or viral load, which likely impacted TB treatment outcomes.[40 46 51] Laboratory values or metabolic biomarkers, such as haemoglobin, haemoglobin A1c or random blood glucose, may also be associated with treatment outcome and worth considering as candidate predictors. There is increasing evidence that diabetes impacts TB treatment outcomes, but caution is warranted about how to best define diabetes in the context of a prediction

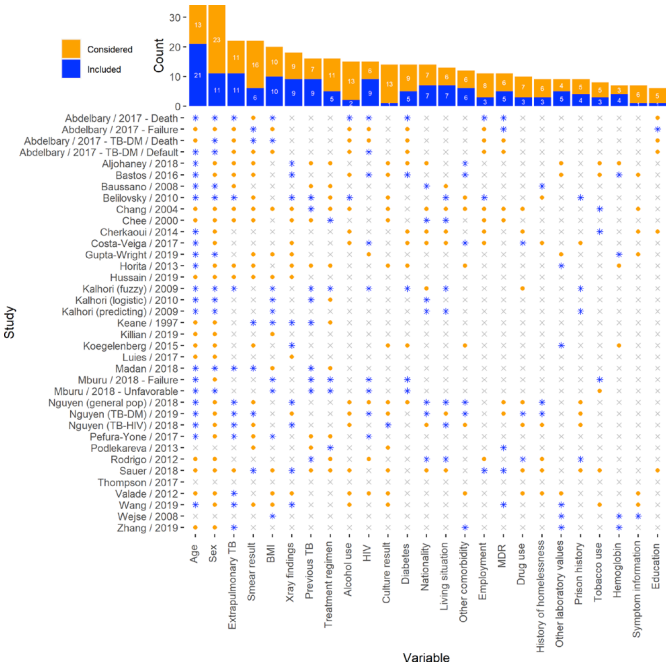

Figure 2 Most common predictors considered and included. Considered: the predictor as evaluated as a candidate predictor prior to multivariable modelling. Included: the predictor was considered and subsequently included in the final multivariable model. BMI, body mass index; MDR, multidrug resistant; TB, tuberculosis.

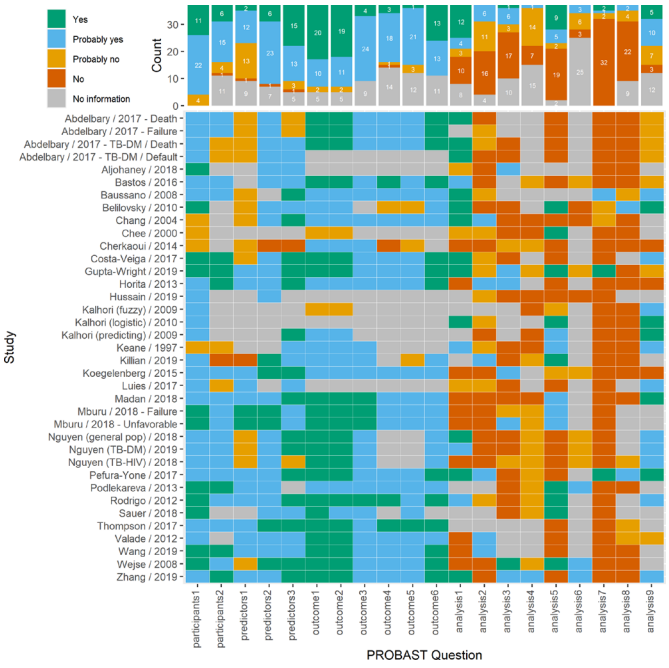

Figure 3 Heatmap of signalling questions from risk of bias assessment with PROBAST. PROBAST questions (additional details in online supplemental file 5) Participants 1: what study design was used and was it appropriate? Participants 2: were all inclusion and exclusion criteria appropriate? Predictors 1: were predictors defined as assessed the same way for all participants? Predictors 2: were predictor assessments made without knowledge of data outcome? Predictors 3: are all predictors available at the time the model was intended to be used? Outcome 1: was the outcome determined appropriately? Outcome 2: was the outcome pre-specified or standard? Outcome 3: were predictors excluded from outcome definition? Outcome 4: was the outcome defined and determined in a similar way for all participants? Outcome 5: was the outcome determined without predictor information? Outcome 6: was the time interval between predictor assessment and outcome determination appropriate? Analysis 1: were there a reasonable number of participants with the outcome? Analysis 2: were continuous and categorical variables handled appropriately? Analysis 3: were all enroled participants included in the analysis? Analysis 4: were participants with missing data handled appropriately? Analysis 5: was selection of predictors based on univariable analysis avoided? Analysis 6: were complexities in data (censoring, competing risks, sampling of control participants) accounted for appropriately? Analysis 7: were relevant model performance measures evaluated appropriately? Analysis 8: were model overfitting, underfitting, and optimism in the model performance accounted for? Analysis 9: do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?.

model to ensure consistency and reproducibility across studies.[59] Behavioural characteristics, such as tobacco use, alcohol use and drug use were rarely included in final prediction models and are difficult to collect objectively, suggesting their role in prediction models for TB treatment outcomes may be limited.

Additionally, several studies excluded participants with HIV, diabetes, extrapulmonary TB or MDR TB, because these factors negatively influence treatment outcomes. However, careful consideration should be given to inclusion/exclusion criteria in prediction model studies, given that information should be available at the time of intended model use, which may not always hold for these aforementioned factors.[60] This is especially questionable for MDR, given that conventional drug-susceptibility testing results are not available for several weeks after TB diagnosis; though more recent advances in rapid molecular methods such as GeneXpert or line-probe assays offer rapid screening.[61]

TB researchers should thoughtfully consider how to appropriately handle complexities of censoring and competing risks in TB outcomes research. Only three studies in this review used survival analysis, despite the long duration of TB treatment outcome assessment and relatively high rates of losses to follow-up across studies, and no studies considered competing risks, such as death due to other causes.[62] Losses to follow-up were frequently excluded, which can lead to selection bias.

Though all included studies were at high risk of bias in the analysis domain, we want to highlight two studies with some exemplary characteristics.[41 44] Pefura-Yone *et al*[41] provide clear explanations of study design, inclusion/exclusion criteria and data collection procedures; TB diagnosis and treatment outcome definitions were standard.[63] Non-linearity of continuous variables was considered with restricted cubic splines, and no continuous variables were categorised or dichotomised; the final

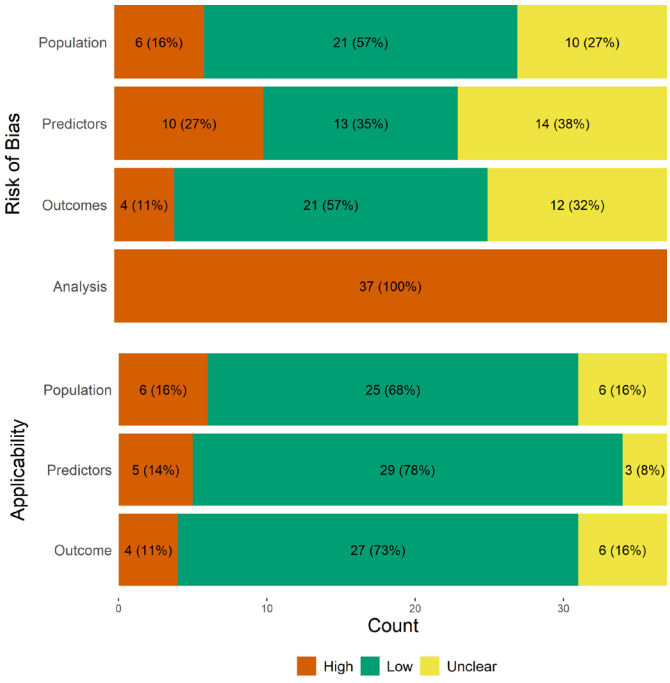

Figure 4 Summary of risk of bias and applicability assessment with PROBAST. PROBAST, Prediction Model Risk of Bias Assessment Tool.

model includes four predictors that are easy to collect and routinely assessed in most TB control programmes, especially those in high-burden settings. The performance of the model was internally validated with bootstrap validation, and the discrimination (c-statistic=0.808) was corrected for optimism. Model calibration was presented graphically with calibration plots. The final model was presented as a nomogram with instructions for use, which facilitates use in external validation studies. Gupta-Wright and colleagues developed and externally validated a clinical risk score to predict mortality in high-burden, low-resource settings.[43] They used clinical trial data with very low amounts of missing data for model development, and externally validated the clinical risk score with data collected independently from two other studies (a clinical trial and a prospective cohort). Given high amounts (42%) of missing data in the validation cohort, they conducted sensitivity analysis using multiple imputation for missing data; the c-statistic differed slightly between complete case and multiple-imputation analyses in the validation cohort (0.68 vs 0.64). Candidate predictors were based on a priori clinical knowledge, previous literature, and required variables were objective, reproducible and available in low-resource settings, consistent with recommended approaches.[26 60 64] Additionally, they reported model performance with the c-statistics and calibration plots for development and validation cohorts, and reported results according to TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guidance.[27 28] Regardless, each of these models requires external validation prior to use in clinical practice.

There are several limitations of this study. Data extraction was subject to reporting in the primary study, which varied widely and was often incomplete, leading to challenges evaluating differences in model performance due to heterogeneities in study populations. Additionally, though most studies reported discrimination, few presented a calibration curve, arguably the most important measure of model performance, further inhibiting assessment and comparison of model performance.[28 65] We did not include external validation studies, which is an essential step for translation to clinical practice. However, several studies in the review did not include the full model equation, which impedes their ability to be externally validated. On searching for studies that externally validated prediction models in this review, we found three studies[66–68] that evaluated the same model (TBscore).[36] Briefly, these studies evaluated the ability of TBscore to monitor treatment response in a new setting,[66] refined the instrument (TBscoreII) using exploratory factor analysis,[67] and then evaluated TBscoreII for use in patients with TB/HIV.[68] To our knowledge, no other studies included in the review were externally validated by other sources. Finally, we excluded 10 studies that were not available in English, Spanish, Portuguese or French; all abstracts were available in English, and none reported model performance metrics, so they likely would have been excluded for different reasons regardless.

The findings of this review not only serve as a comprehensive overview of existing TB outcome prediction models but can act as a resource for future model development and validation of prediction models for TB treatment outcomes. We encourage researchers to focus future TB outcome prediction models on easily collected and readily available predictors that are widely generalisable. We highlight age, sex, extrapulmonary TB, BMI, chest X-ray results, previous TB and HIV as common predictors of TB treatment outcomes. Additionally, when building a new prediction model, it is recommended to first prune the set of considered predictors based on expert opinion and previous literature, rather than univariable analysis or variable selection processes[26 60 64] Future model development or validation studies should adhere to the TRIPOD guidelines, which provide a 22-item checklist and aims to improve the reporting of prediction model development studies.[27 28] We also encourage researchers consider PROBAST criteria to limit bias in design and conduct of prognostic studies.

Prediction models are an important tool in TB management. They can lay the foundation for future impact studies by providing risk estimation to target novel treatment approaches, resource allocation or intensive case management towards patients who are least likely to achieve cure and most likely to benefit from intervention, especially in high-burden and low-resources areas. Use of prediction models can potentially help guide TB treatment practices to achieve the End TB Strategy goal of >90% treatment success, but methodologic rigour and detailed reporting must be improved. Though our

findings suggest that none of the existing models are ready for clinical application without extensive external validation, we hope they direct future researchers to make use of guidelines for development and reporting of prediction models.

Twitter Lauren S. Peetluk @laurenspeetluk

Contributors LP conceptualised the research question, designed the protocol and drafted the manuscript. LP and FR screened studies. FR, PR, DL, VR and TS provided feedback on the research design, original protocol and revised successive drafts of the manuscript. All authors approved the final version of the manuscript.

Funding This work was supported by the National Centre for Advancing Translational Sciences [CTSA Award No. TL1TR000447 to L.S.P.] and the National Institutes of Allergy and Infectious Diseases [F31AI152614-01A1 to L.S.P],. Its contents are solely the responsibility of the authors and do not necessarily represent the official views the National Centre for Advancing Translational Sciences or the National Institutes of Health.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The study protocol is available online at https://osf.io/rz3wp. Most included studies are publicly available. Additional data and code are available upon request.

ORCID iD

Lauren S. Peetluk http://orcid.org/0000-0001-5302-2033

REFERENCES

1 World Health Organization. *Global tuberculosis report 2019*. Geneva, 2019.
2 World Health Organization. *The end TB strategy*. Geneva, 2015.
3 Kerantzas CA, Jacobs WR. Origins of combination therapy for tuberculosis: lessons for future antimicrobial development and application. *mBio* 2017;8:e01586–16.
4 Nahid P, Dorman SE, Alipanah N, *et al*. Official American thoracic Society/Centers for disease control and Prevention/Infectious diseases Society of America clinical practice guidelines: treatment of drug-susceptible tuberculosis. *Clin Infect Dis* 2016;63:e147–95.
5 World Health Organization. *Guildelines for treatment of drug-susceptible tuberculosis and patient care. Licence: CC BY-NC-SA 3.0 IGO*. Geneva: WHO/HTM/TB, 2017.
6 World Health Organization. *Who consolidated guidelines on drug-resistant tuberculosis treatment*. Geneva, 2019.
7 Vasankari T, Holmström P, Ollgren J, *et al*. Risk factors for poor tuberculosis treatment outcome in Finland: a cohort study. *BMC Public Health* 2007;7:1–9.
8 Ramachandran G, Agibothu Kupparam HK, Vedhachalam C, *et al*. Factors influencing tuberculosis treatment outcome in adult patients treated with Thrice-Weekly regimens in India. *Antimicrob Agents Chemother* 2017;61:e02464–16.
9 Abdelbary BE, Garcia-Viveros M, Ramirez-Oropesa H, *et al*. Predicting treatment failure, death and drug resistance using a computed risk score among newly diagnosed TB patients in Tamaulipas, Mexico. *Epidemiol Infect* 2017;145:3020–34.
10 Chaves Torres NM, Quijano Rodríguez JJ, Porras Andrade PS, *et al*. Factors predictive of the success of tuberculosis treatment: a systematic review with meta-analysis. *PLoS One* 2019;14:e0226507.
11 Steyerberg EW, Moons KGM, van der Windt DA, *et al*. Prognosis research strategy (progress) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.
12 Riley RD, Ridley G, Williams K, *et al*. Prognosis research: toward evidence-based results and a Cochrane methods group. *J Clin Epidemiol* 2007;60:863–5.
13 Moons KG, Hooft L, Williams K, *et al*. Implementing systematic reviews of prognosis studies in Cochrane. *Cochrane Database Syst Rev* 2018;10:ED000129.
14 Debray TPA, Damen JAAG, Snell KIE, *et al*. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
15 World Health Organization. *Definitions and reporting framework for tuberculosis - 2013 revision. In: Annex 2, TB case and treatment outcome definitions*. Geneva, 2014.
16 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51–8.
17 Iseman MD. Tuberculosis therapy: past, present and future. *Eur Respir J Suppl* 2002;36:87S–94.
18 Council STSMR. Clinical trial of six-month and four-month regimens of chemotherapy in the treatment of pulmonary tuberculosis: the results up to 30 months. *Tubercle* 1981;62:95–102.
19 Bramer WM, Rethlefsen ML, Kleijnen J, *et al*. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev* 2017;6:1–12.
20 Geersing G-J, Bouwmeester W, Zuithoff P, *et al*. Search filters for finding prognostic and diagnostic prediction studies in MEDLINE to enhance systematic reviews. *PLoS One* 2012;7:3–8.
21 Westgate MJ. revtools: an R package to support article screening for evidence synthesis. *Res Synth Methods* 2019;10:606–14.
22 Innovation VH, Melbourne A. Covidence systematic review sofware. *Covidence* 2016.
23 Harris PA, Taylor R, Minor BL, *et al*. The REDCap Consortium: building an international community of software platform partners. *J Biomed Inform* 2019;95:103208.
24 Harris PA, Taylor R, Thielke R, *et al*. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–81.
25 Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS Med* 2014;11:e1001744.
26 Moons KGM, Wolff RF, Riley RD, *et al*. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019;170:W1–33.
27 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
28 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation* 2015;131:211–9.
29 Cherkaoui I, Sabouni R, Ghali I, *et al*. Treatment default amongst patients with tuberculosis in urban Morocco: predicting and explaining default and post-default sputum smear and drug susceptibility results. *PLoS One* 2014;9:e93574.
30 Keane VP, de Klerk N, Krieng T, *et al*. Risk factors for the development of non-response to first-line treatment for tuberculosis in southern Vietnam. *Int J Epidemiol* 1997;26:1115–20.
31 Chang KC, Leung CC, Tam CM. Risk factors for defaulting from anti-tuberculosis treatment under directly observed treatment in Hong Kong. *Int J Tuberc Lung Dis* 2004;8:1492–8.
32 Chee CB, Boudville IC, Chan SP, *et al*. Patient and disease characteristics, and outcome of treatment defaulters from the Singapore TB control unit--a one-year retrospective survey. *Int J Tuberc Lung Dis* 2000;4:496–503.
33 Luies L, Reenen Mvan, Ronacher K, *et al*. Predicting tuberculosis treatment outcome using metabolomics. *Biomark Med* 2017;11:1057–67.
34 Killian JA, Wilder B, Sharma A. Learning to prescribe interventions for tuberculosis patients using digital adherence data. *Knowledge Discovery And Data Mining* 2019:2430–8.
35 Belilovsky EM, Borisov SE, Cook EF, *et al*. Treatment interruptions among patients with tuberculosis in Russian TB hospitals. *Int J Infect Dis* 2010;14:e698–703.

36 Wejse C, Gustafson P, Nielsen J, *et al*. TBscore: signs and symptoms from tuberculosis patients in a low-resource setting have predictive value and may be used to assess clinical course. *Scand J Infect Dis* 2008;40:111–20.
37 Nguyen DT, Graviss EA. Development and validation of a risk score to predict mortality during TB treatment in patients with TB-diabetes comorbidity. *BMC Infect Dis* 2019;19:10.
38 Kalhori SRN, Zeng X. Fuzzy logic approach to predict the outcome of tuberculosis treatment course destination. *Lecture Notes in Engineering and Computer Science* 2009;2179:774–8.
39 Horita N, Miyazawa N, Yoshiyama T, *et al*. Poor performance status is a strong predictor for death in patients with smear-positive pulmonary TB admitted to two Japanese hospitals. *Trans R Soc Trop Med Hyg* 2013;107:451–6.
40 Koegelenberg CFN, Balkema CA, Jooste Y, *et al*. Validation of a severity-of-illness score in patients with tuberculosis requiring intensive care unit admission. *S Afr Med J* 2015;105:389–92.
41 Pefura-Yone EW, Kuaban C, Assamba-Mpom SA, *et al*. Derivation, validation and comparative performance of a simplified chest X-ray score for assessing the severity and outcome of pulmonary tuberculosis. *Clin Respir J* 2015;9:157–64.
42 Valade S, Raskine L, Aout M, *et al*. Tuberculosis in the intensive care unit: a retrospective descriptive cohort study with determination of a predictive fatality score. *Canadian Journal of Infectious Diseases and Medical Microbiology* 2012;23:173–8.
43 Wang Q, Han W, Niu J, *et al*. Prognostic value of serum macrophage migration inhibitory factor levels in pulmonary tuberculosis. *Respir Res* 2019;20:50.
44 Gupta-Wright A, Corbett EL, Wilson D, *et al*. Risk score for predicting mortality including urine lipoarabinomannan detection in hospital inpatients with HIV-associated tuberculosis in sub-Saharan Africa: derivation and external validation cohort study. *PLoS Med* 2019;16:e1002776.
45 Zhang Z, Xu L, Pang X, *et al*. A clinical scoring model to predict mortality in HIV/TB co-infected patients at end stage of AIDS in China: an observational cohort study. *Biosci Trends* 2019;13:136–44.
46 Podlekareva DN, Grint D, Post FA, *et al*. Health care index score and risk of death following tuberculosis diagnosis in HIV-positive patients. *Int J Tuberc Lung Dis* 2013;17:198–206.
47 Baussano I, Pivetta E, Vizzini L. Predicting tuberculosis treatment outcome in a low-incidence area. *International Journal of Tuberculosis and Lung Disease* 2008;12:1441–8.
48 Costa-Veiga A, Briz T, Nunes C. Unsuccessful treatment in pulmonary tuberculosis: factors and a consequent predictive model. *Eur J Public Health* 2018;28:352–8.
49 Niakan Kalhori SR, Nasehi M, Zeng XJ. A logistic regression model to predict high risk patients to fail in tuberculosis treatment course completion. *IAENG International Journal of Applied Mathematics* 2010;40:1–6.
50 Kalhori SRN, Zeng X-J. PREDICTING THE OUTCOME OF TUBERCULOSIS TREATMENT COURSE IN FRAME OF DOTS - From Demographic Data to Logistic Regression Model. *Proceedings of the International Conference on Health Informatics. SciTePress - Science and and Technology Publications*, 2009:129–34.
51 Madan C, Chopra KK, Satyanarayana S, *et al*. Developing a model to predict unfavourable treatment outcomes in patients with tuberculosis and human immunodeficiency virus co-infection in Delhi, India. *PLoS One* 2018;13:e0204982.
52 Nguyen DT, Jenkins HE, Graviss EA. Prognostic score to predict mortality during TB treatment in TB/HIV co-infected patients. *PLoS One* 2018;13:e0196022–12.
53 Nguyen DT, Graviss EA. Development and validation of a prognostic score to predict tuberculosis mortality. *J Infect* 2018;77:283–90.
54 Pefura-Yone EW, Balkissou AD, Poka-Mayap V, *et al*. Development and validation of a prognostic score during tuberculosis treatment. *BMC Infect Dis* 2017;17:1–9.
55 Rodrigo T, Caylà JA, Casals M, *et al*. A predictive scoring instrument for tuberculosis lost to follow-up outcome. *Respir Res* 2012;13:75.
56 Hussain OA, Junejo KN. Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models. *Inform Health Soc Care* 2019;44:135–51.
57 Sauer CM, Sasson D, Paik KE, *et al*. Feature selection and prediction of treatment failure in tuberculosis. *PLoS One* 2018;13:e0207491–14.
58 Van Wyk SS, Lin HH, Claassens MM. A systematic review of prediction models for prevalent pulmonary tuberculosis in adults. *Int J Tuberc Lung Dis* 2017;21:405–11.
59 Huangfu P, Ugarte-Gil C, Golub J, *et al*. The effects of diabetes on tuberculosis treatment outcomes: an updated systematic review and meta-analysis. *Int J Tuberc Lung Dis* 2019;23:783–96.
60 Steyerberg EW, Models CP, York N. *Clinical prediction models*. New York: Springer, 2009.
61 Sharma SK, Dheda K. What is new in the who consolidated guidelines on drug-resistant tuberculosis treatment? *Indian J Med Res* 2019;149:309–12.
62 Wolbers M, Koller MT, Witteman JCM, *et al*. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 2009;20:555–61.
63 National Tuberculosis Control Program. *Manual for health personnel*. Yaounde, 2012.
64 Royston P, Moons KGM, Altman DG, *et al*. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604.
65 Van Calster B, Nieboer D, Vergouwe Y, *et al*. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:167–76.
66 Janols H, Abate E, Idh J, *et al*. Early treatment response evaluated by a clinical scoring system correlates with the prognosis of pulmonary tuberculosis patients in Ethiopia: a prospective follow-up study. *Scand J Infect Dis* 2012;44:828–34.
67 Rudolf F, Lemvik G, Abate E, *et al*. TBscore II: refining and validating a simple clinical score for treatment monitoring of patients with pulmonary tuberculosis. *Scand J Infect Dis* 2013;45:825–36.
68 Wejse C, Patsche CB, Kühle A, *et al*. Impact of HIV-1, HIV-2, and HIV-1+2 dual infection on the outcome of tuberculosis. *Int J Infect Dis* 2015;32:128–34.
69 Aljohaney A. Mortality of patients hospitalized for active tuberculosis in King Abdulaziz university Hospital, Jeddah, Saudi Arabia. *Saudi Med J* 2018;39:267–72.
70 Bastos HN, Osório NS, Castro AG, *et al*. A prediction rule to stratify mortality risk of patients with pulmonary tuberculosis. *PLoS One* 2016;11:e0162797.
71 Horita N, Miyazawa N, Yoshiyama T, *et al*. Development and validation of a tuberculosis prognostic score for smear-positive in-patients in Japan. *Int J Tuberc Lung Dis* 2013;17:54–60.
72 Mburu JW, Kingwara L, Ester M, *et al*. Use of classification and regression tree (CART), to identify hemoglobin A_{1C} (HbA_{1C}) cut-off thresholds predictive of poor tuberculosis treatment outcomes and associated risk factors. *J Clin Tuberc Other Mycobact Dis* 2018;11:10–16.
73 Thompson EG, Du Y, Malherbe ST, *et al*. Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis* 2017;107:48–58.
74 Kalhori SRN, Zeng X-J. Fuzzy Logic Approach to Predict the Outcome of Tuberculosis Treatment Course Destination. In: *Lecture notes in engineering and computer science*, 2009: 774–8.