

Machine learning-based prediction of tuberculosis recurrence in congregate settings: a retrospective cohort study with temporal external validation

Zijian Fang

Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute)

Nan Jiang

Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute)

Ruoqi Dai

Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute)

Liyun Ai

Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute)

Qingchun Li

Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute)

Qinglin Cheng

`chenghzcdc@sina.com`

Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute)

Research Article

Keywords: tuberculosis recurrence, machine learning, congregate settings, prediction model

Posted Date: June 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9681833/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 Study PopulationTitle page

2 Full title:

3 Machine learning-based prediction of tuberculosis recurrence in congregate settings:
4 a retrospective cohort study with temporal external validation

5 Running title:

6 Prediction of TB recurrence using machine learning

7 Authors:

8 Zijian Fang[†], Nan Jiang[†], Ruoqi Dai, Liyun Ai, Qingchun Li, Qinglin Cheng

9 [†]These authors contributed equally to this work.

10 Affiliations:

11 Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health
12 Supervision Institute), Hangzhou, China

13 Corresponding author:

14 Correspondence: Qinglin Cheng

15 Email: chenghzcdc@sina.com

16 Author contributions:

17 Zijian Fang and Nan Jiang contributed equally to this work and share first authorship.

18 Zijian Fang: conceptualization, data curation, formal analysis, methodology, writing—
19 original draft. Nan Jiang: methodology, validation, writing—review and editing. Ruoqi

20 Dai and Liyun Ai: data collection and data management. Qingchun Li: supervision and
21 project administration. Qinglin Cheng: study design, supervision, funding acquisition,

22 and critical revision of the manuscript. All authors read and approved the final

23 manuscript.

24 Funding:

25 This study was supported by the Zhejiang Provincial Innovative Team for Disease
26 Prevention and Control (No. 2026JKP-03), Zhejiang Province Youth and Youth Work
27 Research Project (No. ZQ2025016), and Hangzhou Medical and Health Science and
28 Technology Project (No. A20263573).

29 Availability of data and materials:

30 The datasets used and/or analyzed during the current study are available from the
31 corresponding author on reasonable request.

32 Ethics approval and consent to participate:

33 This study was approved by the Ethics Committee of the Hangzhou Center for Disease
34 Control and Prevention (Hangzhou Municipal Health Supervision Institute) (Approval
35 No. 2025-EHC-27). The requirement for informed consent to participate was waived
36 by the same committee due to the retrospective nature of the study and the use of
37 de-identified data, in accordance with institutional ethical guidelines and relevant
38 national regulations^[9].

39 Consent for publication:

40 Not applicable.

41 Competing interests:

42 The authors declare that they have no competing interests.

43 Abstract

44 Background: Tuberculosis (TB) recurrence remains a major challenge to TB control,

45 particularly in high-risk congregate settings, where tools for individualized recurrence
46 prediction using routinely collected data remain limited.

47 Methods: We conducted a retrospective cohort study including 5,994 individuals with
48 pulmonary TB in congregate settings in Hangzhou, China, from January 2005 to March
49 2024. The outcome was TB recurrence within one year after treatment completion.

50 Participants were temporally divided into a training cohort (January 2005–December
51 2022; n=5,407) and a validation cohort (January 2023–March 2024; n=587).

52 Participants with missing values in key predictors were excluded using complete-case
53 analysis. Feature selection was performed using LASSO regression. Six predictive
54 models were developed and compared. Model performance was evaluated using
55 discrimination (AUC), calibration (Brier score and calibration plots), and clinical utility
56 through decision curve analysis.

57 Results: XGBoost demonstrated the best overall performance. In the training cohort,
58 the model achieved an AUC of 0.943 (95% CI: 0.935–0.952), sensitivity of 0.476,
59 specificity of 0.988, and Brier score of 0.045. In the validation cohort, the AUC was
60 0.879 (95% CI: 0.803–0.955), with sensitivity of 0.455, specificity of 0.967, and Brier
61 score of 0.071. Key predictors included household registration, age, imaging findings,
62 fixed-dose combination therapy, and delay in healthcare seeking. Decision curve
63 analysis demonstrated superior net benefit across clinically relevant threshold
64 probabilities.

65 Conclusions: This machine learning–based model demonstrated strong predictive
66 performance for TB recurrence in congregate settings and may facilitate precision

public health by enabling risk-stratified follow-up, optimized surveillance intensity, and targeted secondary prevention strategies in high-risk populations.

Keywords: tuberculosis recurrence; machine learning; congregate settings; prediction model

Post-treatment tuberculosis recurrence (hereafter referred to as TB recurrence) is defined as the re-emergence of active TB disease in individuals who have previously completed a full course of anti-TB treatment and were classified as cured. This condition encompasses both relapse, resulting from reactivation of the original infection, and reinfection with a new strain of *Mycobacterium tuberculosis*^[1].

Due to the lack of molecular genotyping data in most routine surveillance systems, these two entities cannot be reliably distinguished in epidemiological studies and are therefore commonly analyzed as a composite outcome. TB recurrence remains a major challenge to global TB control efforts, as affected individuals often present with more severe clinical manifestations, including extensive pulmonary destruction, cavitary disease, and a higher likelihood of drug resistance^[2,3].

These cases frequently require prolonged treatment with second-line drugs, which are associated with increased toxicity, higher healthcare costs, and poorer outcomes, including elevated morbidity and mortality^[4, 5]. In addition to clinical burden, TB recurrence is frequently accompanied by substantial psychosocial stress, including stigma, social isolation, and financial hardship, which may further compromise treatment adherence and quality of life. ^[6, 7]

TB is transmitted primarily through airborne spread of *Mycobacterium tuberculosis*,

particularly in settings characterized by prolonged close contact, crowding, and suboptimal ventilation. Although infection prevention and control (IPC) efforts have traditionally focused on healthcare facilities, growing evidence indicates that non-healthcare institutional environments play a critical role in sustaining TB transmission. According to the World Health Organization (WHO) Operational Handbook on Tuberculosis Infection Prevention and Control^[8], congregate settings are defined as institutional (non-healthcare) environments where individuals reside or work in close proximity, including prisons, refugee camps, army barracks, shelters, dormitories, hospices, and nursing homes. Such environments often present structural and operational characteristics—shared accommodation, high population density, delayed case detection, and variable healthcare access—that facilitate airborne transmission and increase TB incidence compared with the general population. However, predictive models for TB recurrence in high-risk congregate settings remain scarce, and the application of machine learning approaches in such contexts is still limited.

Although previous studies have investigated predictors for TB recurrence, most have focused on general populations or specific clinical cohorts, with limited attention to high-risk congregate settings. Moreover, few studies have applied machine learning approaches to develop individualized risk prediction models in such environments.

This study aimed to develop and validate a machine learning-based model for predicting tuberculosis (TB) recurrence in congregate settings. By leveraging routinely collected clinical and demographic data, the model seeks to enable early identification of high-risk individuals and facilitate timely preventive interventions.

1. Materials and Methods

This retrospective cohort study was reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.

1.1 Study Population

This retrospective study included 5,994 individuals from congregate settings who were diagnosed with pulmonary tuberculosis (TB) and registered for treatment in Hangzhou, China, between January 2005 to March 2024.

Inclusion criteria:

(1) Patients who completed the full course of anti-tuberculosis treatment and were classified as cured; (2) Individuals who were in congregate settings at the time of initial diagnosis.

Exclusion criteria:

(1) Patients with tuberculosis (TB) whose treatment outcomes could not be classified; (2) Patients who died during the study period; (3) Patients who were still receiving anti-tuberculosis treatment; (4) Patients with missing data in key predictors were excluded (complete-case analysis). The proportion of excluded cases due to missing data was 1.2%(n=66).

This study was approved by the Ethics Committee of the Hangzhou Center for Disease Control and Prevention (Hangzhou Municipal Health Supervision Institute) (Approval No. 2025-EHC-27). The requirement for informed consent to participate was waived by the same committee due to the retrospective nature of the study and the use of

de-identified data, in accordance with institutional ethical guidelines and relevant national regulations^[9].

A total of 5,407 confirmed TB cases from congregate settings who completed the full course of anti-tuberculosis treatment and were classified as cured between January 2005 and Dec 2022 were included in the training set, while 587 cases meeting the same criteria between January 2023 to March 2024 were used as the validation set. Although no formal a priori sample size calculation was performed, all eligible participants during the study period were consecutively included. The number of TB recurrence events exceeded recommended events-per-variable thresholds, supporting model stability and minimizing overfitting.

In this study, congregate settings were operationally defined as correctional facilities (prisons), hospices, dormitories, nursing homes, and workplace-based residential environments, including factories and mining sites. These environments share key epidemiological features conducive to TB transmission, including prolonged interpersonal contact, shared living spaces, and constrained ventilation conditions.

Military barracks, refugee camps, and homeless shelters-also included within the WHO conceptual framework-were not represented in the available surveillance dataset. Consequently, the study population represents a defined subset of WHO-described congregate environments rather than the full spectrum of institutional settings.

1.2 Data Collection

Clinical and demographic data were extracted from the electronic medical record systems of 14 designated tuberculosis hospitals in Hangzhou, China. The collected

variables included age, sex, household registration status, current residence, presence of symptoms, microbiological diagnosis, drug resistance, imaging findings, treatment regimen, method of case detection, managing institution, use of fixed-dose combination (FDC) therapy, HIV (human immunodeficiency virus) status, and recurrence of TB within one year.

In this study, the primary outcome was post-treatment TB recurrence, defined as the occurrence of active TB disease within one year after completion of anti-tuberculosis treatment among individuals initially classified as cured. Owing to the absence of molecular genotyping data, recurrence events could not be differentiated into relapse or reinfection and were therefore analyzed as a composite endpoint.

All predictors included in the model were defined based on information available at or before treatment initiation to minimize potential data leakage.

All data were collected by trained personnel from the Hangzhou CDC and Health Supervision Authority. “Symptoms” referred to the presence of typical TB manifestations at the time of initial diagnosis, such as persistent cough or sputum for ≥ 2 weeks, hemoptysis, chest tightness, low-grade fever, night sweats, fatigue, loss of appetite, or weight loss.

Microbiological diagnosis was categorized as positive, negative, or not tested. A positive result included at least one of the following: smear-positive sputum, positive culture, or molecular confirmation of *Mycobacterium tuberculosis*.

Drug resistance referred to resistance to any first-line or second-line anti-TB drug.

Imaging findings were based on chest radiography or CT at initial diagnosis and

classified as abnormal or normal.

Treatment regimens included individualized anti-TB treatment regimens (ITR) and non-individualized treatment regimens (NITR). ITRs were tailored to each patient based on clinical, microbiological, and sociodemographic characteristics, as opposed to standard protocol-based treatment.

The method of case detection was classified as active detection (e.g., screening of high-risk groups or in high-prevalence areas) or passive detection (e.g., symptomatic presentation or incidental detection during physical examination).

The managing institution referred to the facility responsible for the initial diagnosis and treatment of the patient, classified as either county-level or higher-level designated TB hospitals.

FDCs were defined as fixed-dose combinations containing two or more anti-TB drugs in specified ratios. HIV co-infection was defined as confirmed HIV infection at the time of initial TB diagnosis.

1.3 Statistical Analysis

All statistical analyses were conducted using R software version 4.4.1. Continuous variables with non-normal distribution were expressed as medians with interquartile ranges (IQR), and group comparisons were performed using the Wilcoxon rank-sum test. Categorical variables were expressed as proportions and compared using the chi-square test.

Predictive models for tuberculosis recurrence were developed using six machine learning algorithms: logistic regression (LR), random forest (RF), Gaussian naïve Bayes

(GNB), decision tree (DT), support vector machine (SVM), and extreme gradient boosting (XGBoost). Participants were temporally divided into a development cohort and an independent temporal validation cohort based on date of treatment completion to preserve chronological separation. All preprocessing procedures, including feature selection, were performed exclusively within the development cohort to avoid data leakage. For XGBoost, a binary logistic objective function was used with nrounds set to 100, while other key hyperparameters (max depth, eta, subsample, and colsample bytree) were retained at default settings. To ensure methodological comparability across the six machine learning models, a standardized parameter framework was applied; however, this approach may have limited full optimization of individual model performance. Model performance was subsequently evaluated in the temporal validation cohort.

The dataset was temporally divided into a development cohort (January 2005–December 2022) and a temporal validation cohort (January 2023–March 2024), preserving chronological separation to better evaluate model generalizability, with the split stratified by TB recurrence to preserve the distribution of events. Missing data were handled using complete-case analysis; participants with missing values in any predictor or outcome were excluded. Feature selection was performed in the training set using least absolute shrinkage and selection operator (LASSO) regression to identify the most informative predictors.

Calibration of predicted probabilities was assessed using calibration plots, in which predicted probabilities were grouped into deciles and plotted against observed event

rates. Brier scores were also calculated to quantify the overall agreement between predicted probabilities and observed outcomes. ROC curves and calibration plots were generated and saved in PNG format for visualization. Decision curve analysis was subsequently conducted to evaluate the clinical usefulness of each model by calculating the net benefit across a range of threshold probabilities. Net benefit was defined as the difference between the proportion of true positives and the proportion of false positives, weighted by the relative harm of false-positive and false-negative results.

To enhance model interpretability, SHapley Additive exPlanations (SHAP) values were calculated using the `iml` R package. A Predictor object was constructed to link model probability outputs with SHAP values at the individual level. Variable importance was extracted and visualized using bar plots generated with `ggplot2` to illustrate the marginal contribution of each feature to the probability of TB recurrence.

All analyses were conducted using the `caret`, `pROC`, `randomForest`, `e1071`, `xgboost`, `rpart`, `dplyr`, `ggplot2`, and `iml` packages. A significance level of $p < 0.05$ was considered statistically significant.

2. Results

2.1 Comparison of Baseline Characteristics Between Training and Validation Sets

A total of 5,994 individuals in congregate settings with pulmonary tuberculosis were included in the study, with 5,407 cases (90.21%) assigned to the training set and 587 cases (9.79%) to the validation set. The demographic characteristics of patients in the two datasets are presented in Table 1.

Table 1. Baseline characteristics of individuals with pulmonary tuberculosis in
congregate settings in the training and validation cohorts

Variable	Category	Training set (n = 5407)	Validation set (n = 587)	Z / χ^2 value	P- value
Age(years)	M (IQR)	30 (24.0,39.0)	31 (25.0,39.0)	-0.966	0.334
Sex	Male	5133 (94.9%)	546 (93.0%)	3.909	0.048
	Female	274 (5.1%)	41 (7.0%)		
Household	Local	3319 (61.4%)	355 (60.5%)	0.183	0.668
registration	Non-local	2088 (38.6%)	232 (39.5%)		
Current	Urban	1411 (26.1%)	163 (27.8%)	0.765	0.382
residence	Rural	3996 (73.9%)	424 (72.2%)		

Note: Age is presented as Median (IQR); group comparisons were made using the Wilcoxon rank-sum test for continuous variables and the chi-square test for categorical variables.

2.2 Feature Selection Using LASSO Regression

The least absolute shrinkage and selection operator (LASSO) regression was employed to identify the most relevant predictors associated with TB recurrence. LASSO is a regularization technique that performs variable selection and coefficient shrinkage simultaneously by imposing an L1 penalty, thereby reducing model complexity and minimizing the risk of overfitting.

All candidate variables, including demographic characteristics, clinical features, and treatment-related factors, were incorporated into the model. Categorical variables

were transformed into dummy variables using a model matrix prior to analysis. A 10-fold cross-validation procedure was conducted to determine the optimal penalty parameter (λ). Two candidate lambda values were considered: $\lambda_{\min}$ (minimum cross-validated error) and λ_{1se} (most regularized model within one standard error). Two candidate values were considered: the λ corresponding to the minimum mean cross-validated error ($\lambda_{\min}$) and the λ within one standard error of the minimum (λ_{1se}), which provides a more parsimonious model.

In this study, the optimal λ values were identified as $\lambda_{\min} = 0.00067$ and $\lambda_{1se} = 0.01093$. At $\lambda_{\min}$, 14 predictors with non-zero coefficients (excluding the intercept) were selected, whereas only 7 predictors were retained at λ_{1se} , representing a substantial reduction in model complexity. The λ_{1se} model yielded a more parsimonious and stable subset of variables, while $\lambda_{\min}$ was used for primary feature selection in subsequent analyses.

Furthermore, several predictors were consistently identified across both $\lambda_{\min}$ and λ_{1se} models, indicating the robustness of the feature selection process. The variables selected by LASSO regression were subsequently incorporated into downstream analyses, including multivariable logistic regression and machine learning models, to evaluate their predictive performance.

The LASSO coefficient path and cross-validation results are shown in Figure 1.

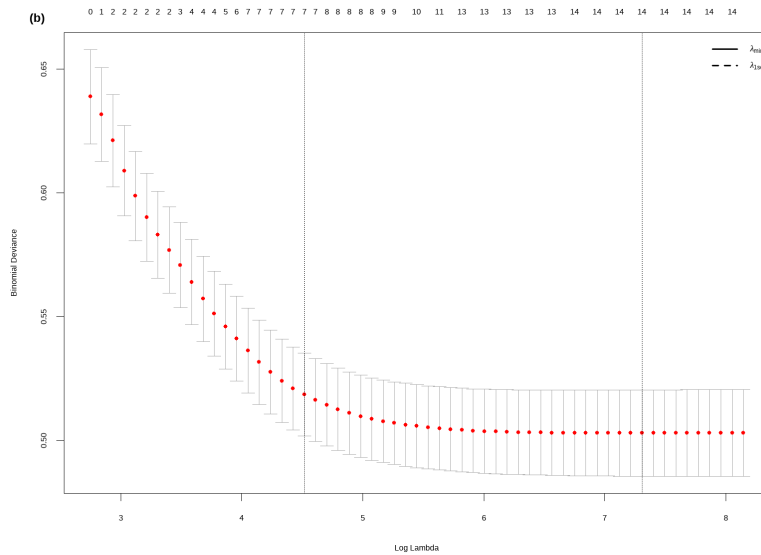
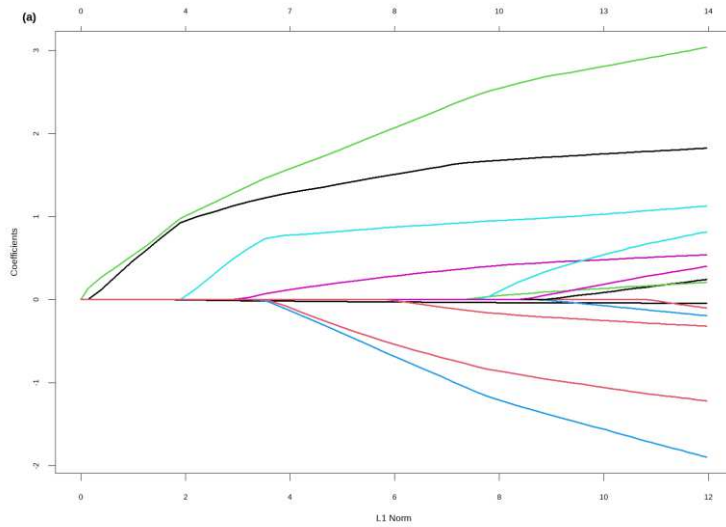


Figure 1. LASSO regression results showing (a) coefficient profiles of the features and (b) 10-fold cross-validation for selecting the optimal lambda value.

(a) Coefficient curves for clinical features based on LASSO regression. Each curve represents the trajectory of a feature's coefficient as the penalty parameter ($\log_{\lambda}$) changes. (b) Screening of optimal variables by 10-fold cross-validation. The two vertical dotted lines indicate the optimal values of the penalty parameter λ : the left corresponds to $\lambda_{\min}$ ($\lambda = 0.00067$), representing the minimum mean cross-validated error, and the right corresponds to λ_{1se} , representing the most regularized model

within one standard error of the minimum. At $\lambda_{\min}$, 14 features with non-zero coefficients were selected.

2.3 Model Performance Comparison

We developed six machine learning (ML) models to assess the risk of tuberculosis (TB) recurrence among individuals in congregate settings, including logistic regression (LR), random forest (RF), Gaussian naïve Bayes (GNB), decision tree (DT), extreme gradient boosting (XGBoost), and support vector machine (SVM).

In the training set, the XGBoost model demonstrated the best overall performance, achieving the highest predictive accuracy (AUC = 0.943, 95% CI: 0.935–0.952), with a sensitivity of 0.476, specificity of 0.988, and F1 score of 0.601. The combination of relatively low sensitivity and high specificity indicates that the model tends to minimize false-positive classifications, potentially at the expense of missing a proportion of individuals with higher predicted probabilities of recurrence.

Similarly, in the validation set, XGBoost maintained superior performance, with an AUC of 0.879 (95% CI: 0.803–0.955), sensitivity of 0.455, specificity of 0.967, and F1 score of 0.541, supporting its robustness and generalizability.

The relatively low sensitivity observed in the XGBoost model suggests a conservative classification strategy prioritizing the reduction of false positives, which may be appropriate in settings where unnecessary interventions carry substantial costs.

Detailed performance metrics of all six models are summarized in Table 2. The receiver operating characteristic (ROC) curves for the models are shown in Figure 2.

Table 2. Performance comparison of six machine learning models for predicting TB

Dataset	Model	AUC (95% CI)	Sensitivity	Specificity	F1 Score	Brier score
Training	LR	0.810 (0.793– 0.827)	0.178	0.993	0.286	0.075
	RF	0.900 (0.883– 0.916)	0.622	0.983	0.699	0.067
	GNB	0.821 (0.805– 0.838)	0.355	0.943	0.377	0.100
	DT	0.729 (0.707– 0.751)	0.226	0.990	0.342	0.076
	XGBoost	0.943 (0.935– 0.952)	0.476	0.988	0.601	0.045
	SVM	0.847 (0.832– 0.863)	0.302	0.994	0.446	0.066
Validation	LR	0.835 (0.760– 0.910)	0.227	0.961	0.303	0.095
	RF	0.834 (0.734– 0.935)	0.273	0.980	0.387	0.090
	GNB	0.779 (0.655– 0.904)	0.864	0.549	0.345	0.404

DT	0.839 (0.764– 0.913)	0.227	0.980	0.333	0.092
XGBoost	0.879 (0.803– 0.955)	0.455	0.967	0.541	0.071
SVM	0.871 (0.809– 0.932)	0.227	0.987	0.345	0.091

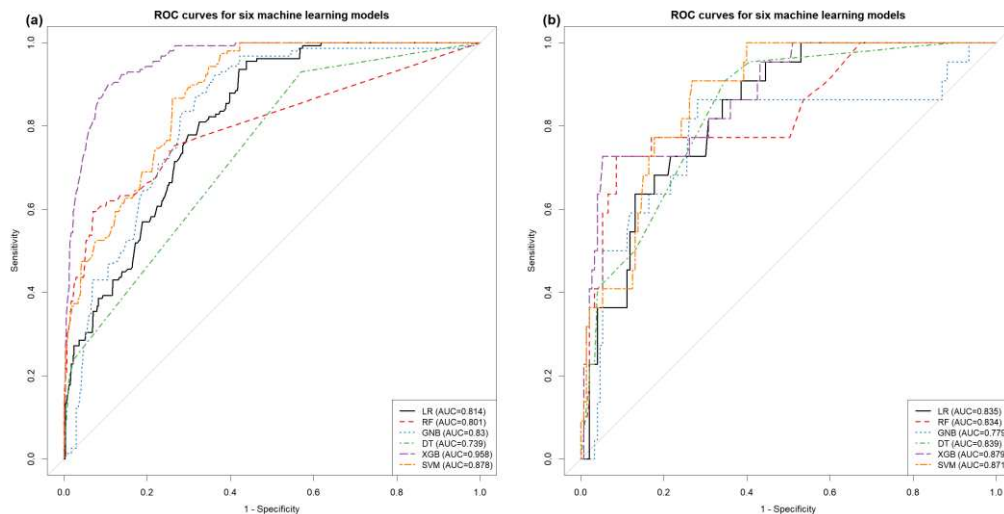


Figure 2. ROC curves for six machine learning models in (a) the training set and (b)

the validation set. The XGBoost model achieved the highest AUC in both datasets.

Calibration of predicted probabilities was assessed using decile-based calibration

plots and Brier scores. Figure 3 shows the calibration curves for the training set,

while Figure 4 displays the calibration curves for the validation set. Predicted

probabilities were grouped into ten bins and plotted against the observed TB

recurrence rates. The red dashed line represents perfect calibration (45-degree line).

XGBoost demonstrated the best calibration on the validation set (Brier = 0.071), with

predicted probabilities closely following the ideal 45-degree line. Random forest

(0.067) and SVM (0.066) also exhibited good agreement between predicted and observed TB recurrence probabilities. Logistic regression (0.075) and decision tree (0.076) showed moderate calibration, and Gaussian naïve Bayes performed the worst (0.100). These results indicate that XGBoost achieved the highest discrimination in both the training (AUC = 0.943) and validation cohorts (AUC = 0.879).

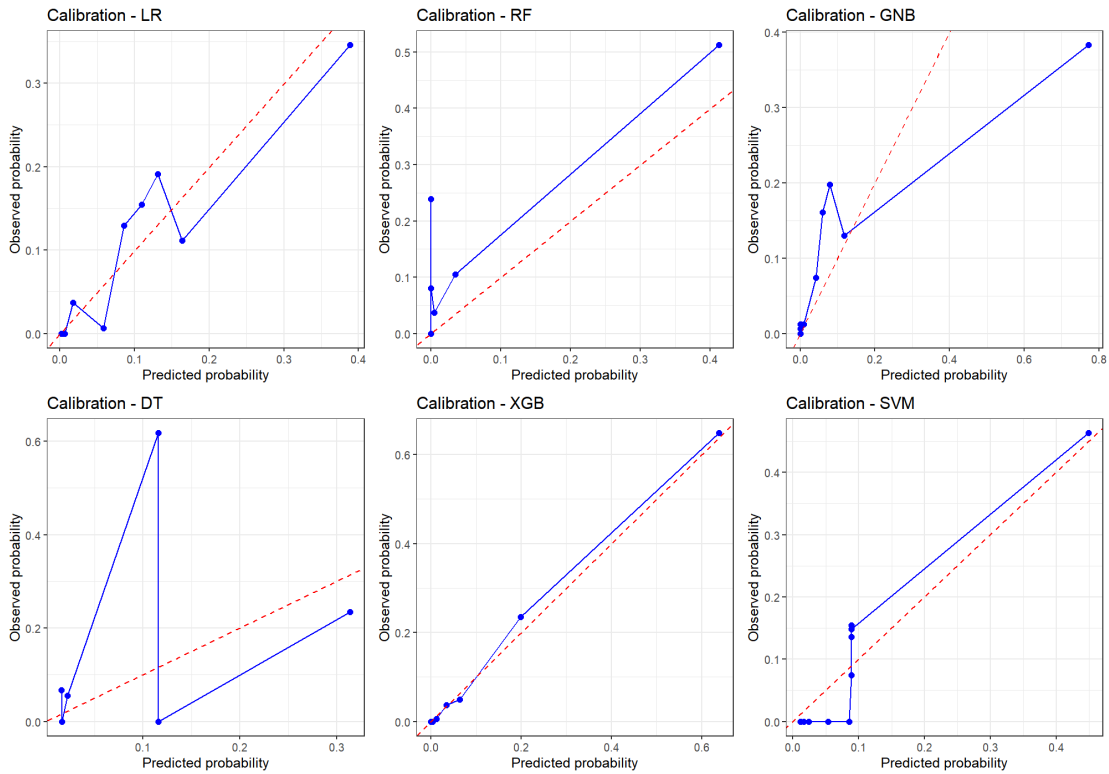


Figure 3. Calibration curves of the six predictive models in the training set

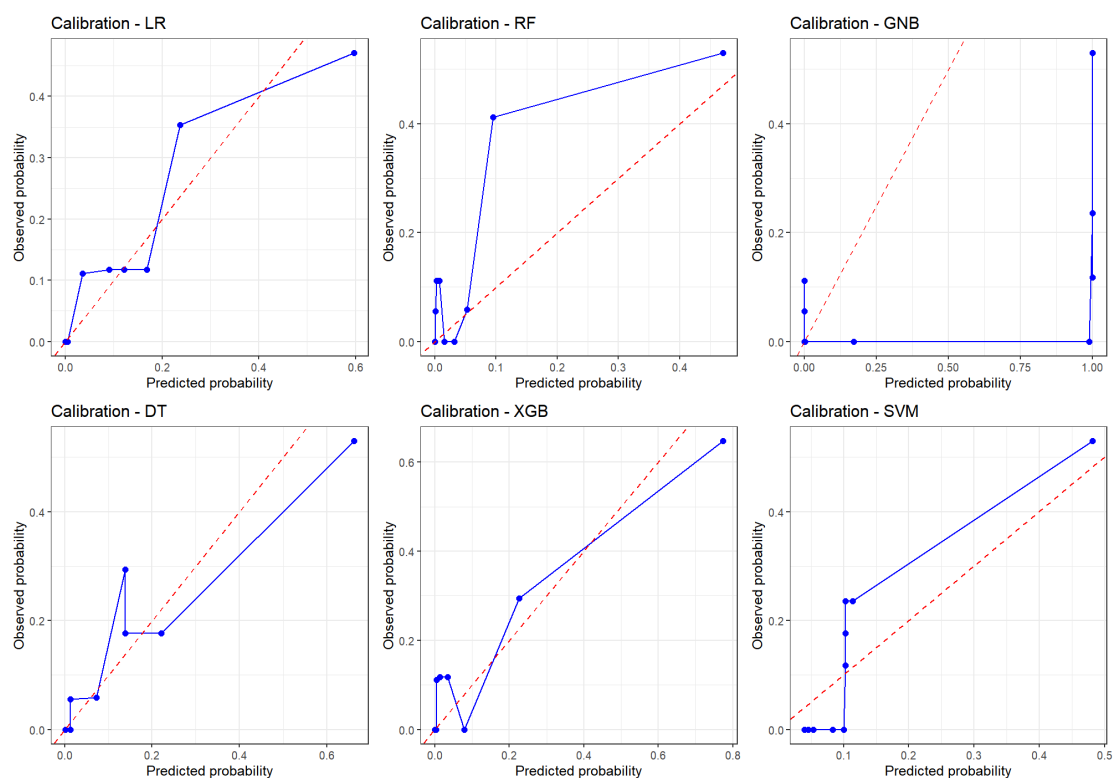


Figure 4. Calibration curves of the six predictive models in the validation set

2.4 Feature Importance Based on SHAP Values in the XGBoost Model

To improve model interpretability, SHAP (SHapley Additive exPlanations) was used to quantify the contribution of each feature to the model output.

The analysis showed that the most influential features contributing to the model-predicted probability of TB recurrence were, in descending order of importance:

household registration, age, imaging findings, use of fixed-dose combination (FDC) therapy, delay in seeking care, current residence, treatment regimen, microbiological diagnosis, presence of symptoms, method of case detection, sex, managing institution, and drug resistance. (Figure 5).

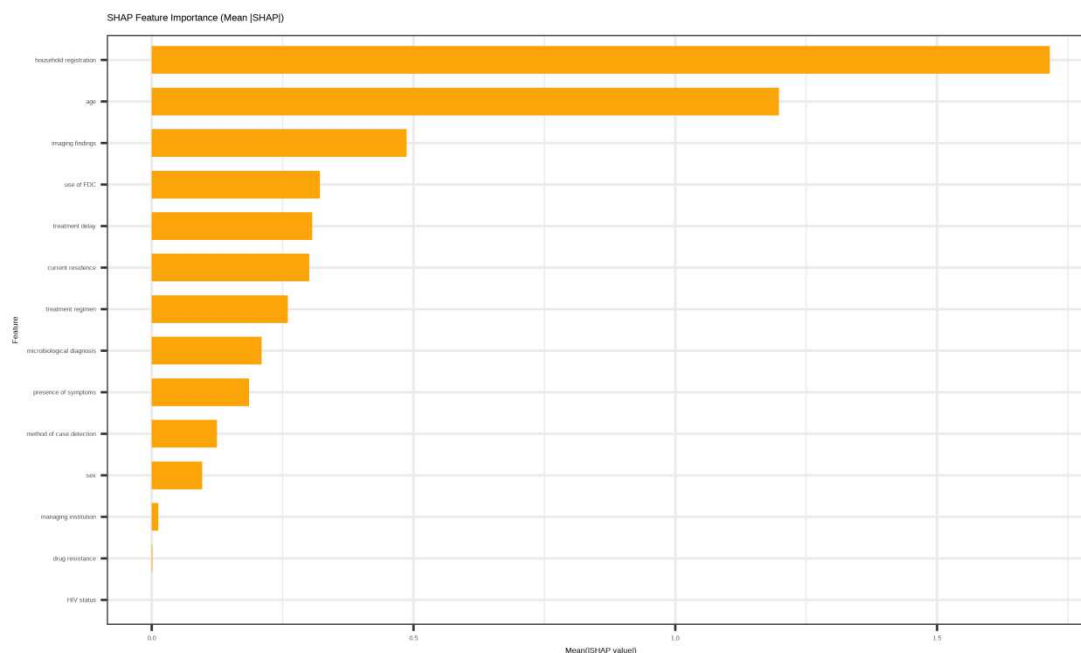


Figure 5. Feature Importance Ranking Based on Mean Absolute SHAP Values

2.5 SHAP Summary Plot of Predictor Contributions

The SHAP summary plot provided an integrated view of feature importance and directional effects on the model-predicted probability of tuberculosis recurrence. Household registration emerged as the most influential predictor, with non-local status contributing substantially to higher model-predicted probability of TB recurrence. Age exhibited heterogeneous effects across the prediction spectrum.

Clinical indicators reflecting disease severity, including abnormal imaging findings, bacteriological positivity, and the presence of symptoms, were associated with higher SHAP values, indicating positive contributions to the model-predicted probability of TB recurrence. Delayed healthcare seeking also showed a strong positive contribution to the model-predicted probability, potentially reflecting prolonged disease progression and barriers to timely care.

The observed association between fixed-dose combination (FDC) use and higher

model-predicted probability of TB recurrence should be interpreted cautiously as a non-causal signal potentially influenced by confounding by indication.

Conversely, rural residence, individualized treatment regimens, and active case detection were associated with lower SHAP values, suggesting negative contributions to the model-predicted probability of TB recurrence within the model context. Other variables, including sex, managing institution, drug resistance, and HIV status, showed relatively limited contributions. (Figure 6).

Overall, the SHAP summary plot indicates that the model's predictions are jointly driven by socio-demographic characteristics, clinical features, and healthcare-related variables, reflecting a multidimensional prediction structure.

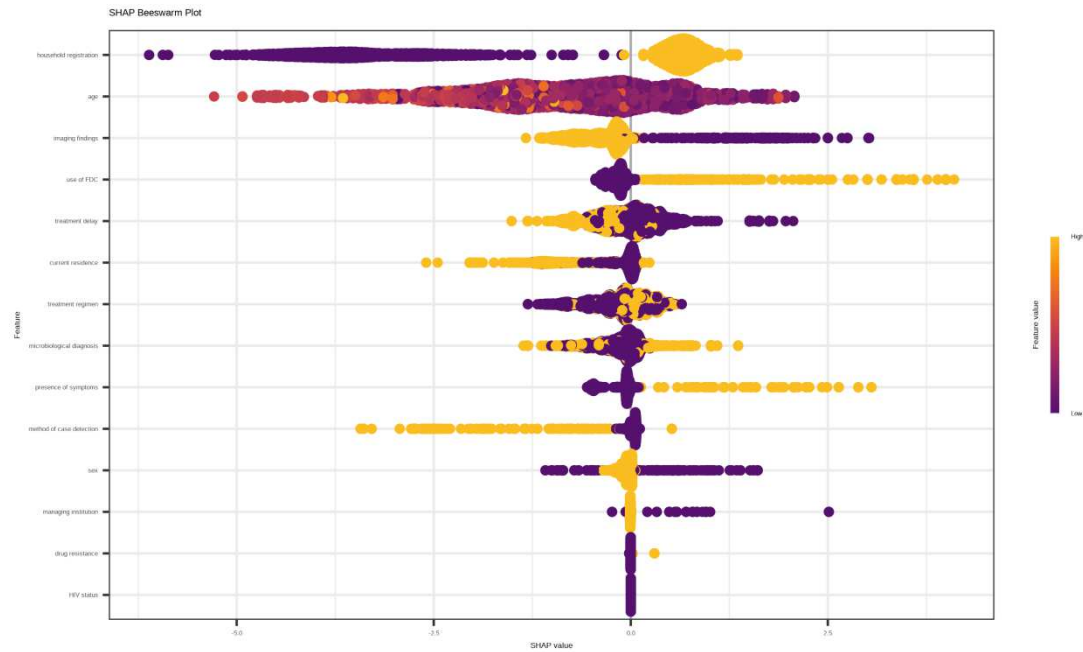


Figure 6. SHAP summary plot illustrating feature effects on TB recurrence prediction

2.6 SHAP Dependence Analysis of Age

Age exhibited a clear non-linear relationship with the model-predicted probability of TB recurrence, characterized by distinct risk strata. Individuals aged approximately 20-

30 years exhibited relatively higher SHAP values, indicating higher model-predicted probabilities of TB recurrence. In contrast, those in the middle-aged group (approximately 40-55 years) showed markedly lower SHAP values, indicating lower contributions to the model-predicted probability of TB recurrence. Among older individuals (>60 years), SHAP values tended to increase again, although with substantial dispersion, suggesting higher model-predicted probabilities, with substantial variability accompanied by pronounced heterogeneity. (Figure 7).

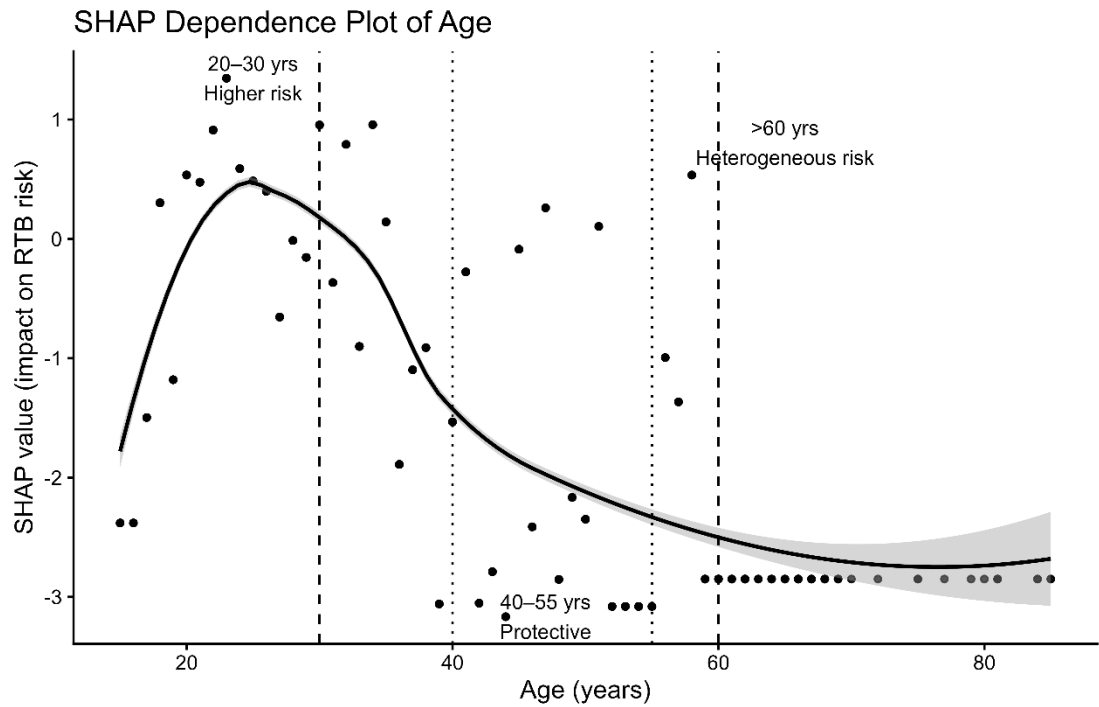


Figure 7. SHAP dependence plot illustrating the effect of age on TB recurrence risk.

2.7 Decision Curve Analysis (DCA)

To further evaluate the clinical utility of the predictive model, decision curve analysis (DCA) was performed for both the training and validation datasets (Figure 8).

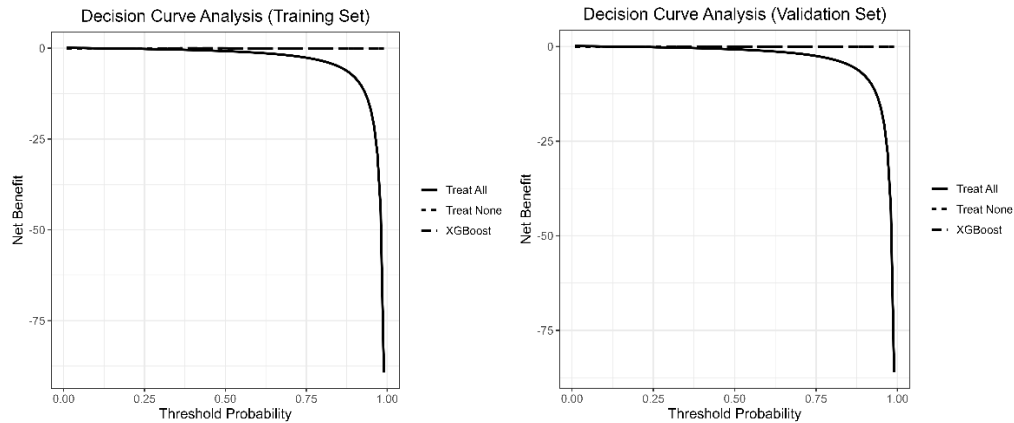


Figure 8. Decision curve analysis (DCA) of the XGBoost model for Decision curve analysis of the XGBoost model for TB recurrence in the training set (a) and validation set (b).

Decision curve analysis demonstrated that the XGBoost model provided consistently greater net benefit than both “treat-all” and “treat-none” strategies across a wide range of threshold probabilities in the training set, particularly between 0.05 and 0.60.

In the temporal validation set, the model maintained a favorable net benefit profile within a threshold range of approximately 0.10 to 0.55. These findings indicate that applying the model to guide risk-based follow-up or targeted preventive interventions may achieve a more favorable balance between true-positive identification and unnecessary interventions compared with default strategies.

The consistency of these results across datasets supports the potential clinical utility and generalizability of the model in real-world settings.

3. Discussion

This study developed and temporally validated a machine learning–based model for predicting tuberculosis recurrence in congregate settings. The model demonstrated

strong discrimination and consistent performance across datasets, supporting its robustness and potential applicability in real-world settings.

To our knowledge, this is one of the few studies to specifically focus on TB recurrence prediction in congregate environments using machine learning approaches and routinely collected surveillance data.

A key strength of this study is its focus on congregate settings, which represent high-risk yet under-investigated contexts in TB epidemiology. Unlike prior studies^[10] conducted in general populations or specialized clinical cohorts, this study captures the structural and operational characteristics of institutional environments, including high population density, prolonged close contact, and constrained healthcare access, all of which are known to facilitate TB transmission^[3]. The consistent performance observed in the external validation cohort further supports the robustness and potential applicability of the model in similar real-world settings.

Several predictors identified by the model are clinically and biologically plausible. Indicators of baseline disease severity, including abnormal imaging findings, bacteriological positivity, and the presence of symptoms, were associated with higher model-predicted probability of TB recurrence, likely reflecting increased mycobacterial burden and more extensive pulmonary involvement. These findings are consistent with previous studies reporting that more severe baseline disease is associated with an increased risk of tuberculosis recurrence^[2,11]; however, in the present study, these variables reflect their contribution to model-predicted probability rather than causal effects. Delayed healthcare seeking also contributed

418 substantially to model predictions, potentially capturing both prolonged infectious
419 periods and systemic barriers to care.

420 Socio-demographic characteristics played a prominent role. Non-local household
421 registration was associated with higher predicted risk, consistent with previous
422 evidence linking population mobility to fragmented care, treatment interruption, and
423 increased recurrence risk ^[12]. In congregate environments, these vulnerabilities may
424 be amplified by structural constraints. Conversely, individualized treatment regimens,
425 active case detection, and rural residence were associated with lower predicted risk,
426 potentially reflecting improved continuity of care and earlier disease identification.

427 The association between fixed-dose combination therapy and increased predicted
428 risk should be interpreted cautiously, as it may reflect confounding by indication
429 rather than a causal effect. Similar observations have been reported in previous
430 studies ^[13], where treatment allocation was influenced by disease severity and
431 clinical characteristics. Therefore, this finding should not be interpreted as evidence
432 against FDC use but rather as a reflection of underlying clinical complexity.

433 From a methodological perspective, the superior performance of XGBoost likely
434 reflects its ability to capture complex non-linear relationships and higher-order
435 interactions among predictors. This was particularly evident in the non-linear
436 association between age and recurrence risk, which followed a U-shaped pattern.

437 This pattern may reflect behavioral factors in younger individuals and increased
438 vulnerability among older adults due to comorbidities or immune decline.

Importantly, decision curve analysis demonstrated that the model provides meaningful clinical benefit across a range of threshold probabilities, supporting its potential use in guiding risk-stratified follow-up and targeted interventions. Such an approach may improve resource allocation efficiency in high-burden, resource-limited settings, as highlighted in global TB control strategies.

Several limitations should be acknowledged. First, the retrospective design may introduce information bias and limit the availability of important variables such as treatment adherence and behavioral factors. Second, due to the absence of molecular genotyping data, relapse and reinfection could not be distinguished; however, the composite outcome reflects real-world surveillance conditions and aligns with programmatic decision-making needs. Third, the study population represents a subset of congregate environments, and caution is warranted when generalizing findings to other settings. Fourth, the relatively low event rate and class imbalance may have influenced model sensitivity, despite the strong overall discrimination observed. This highlights the potential need for further validation in larger cohorts or settings with higher TB recurrence rates.

In conclusion, this machine learning–based model using routinely collected surveillance data demonstrated strong discrimination and acceptable calibration for predicting tuberculosis recurrence in congregate settings. Beyond individual-level prediction, the model has potential public health utility: it may guide risk-stratified follow-up, optimize allocation of limited surveillance resources, and support targeted secondary prevention interventions in high-risk congregate populations.

References

- [1] Dearden E, van Leth F, Molemans M, et al. The role of neighborhood factors in the cumulative number of episodes of recurrent tuberculosis in Cape Town. *J Clin Tuberc Other Mycobact Dis.* 2026; 43: 100591. doi: 10.1016/j.jctube.2026.100591.
- [2] Nagu TJ, Mboka MA, Nkrumbih ZF, et al. Clinical and imaging features of adults with recurrent pulmonary tuberculosis: a prospective case-controlled study. *Int J Infect Dis.* 2021; 113: S33–S39. doi: 10.1016/j.ijid.2021.09.001.
- [3] World Health Organization. Global tuberculosis report 2023. Geneva: WHO; 2023. Available from: <https://www.who.int/publications/i/item/9789240083851>. Accessed 8 May 2026.
- [4] Akalu TY, Clements ACA, Wolde HF, et al. Economic burden of multidrug-resistant tuberculosis on patients and households: a global systematic review and meta-analysis. *Sci Rep.* 2023; 13(1): 22361. doi: 10.1038/s41598-023-49555-8.
- [5] Chung-Delgado K, Guillen-Bravo S, Revilla-Montag A, et al. Mortality among MDR-TB cases: comparison with drug-susceptible tuberculosis and associated factors. *PLoS One.* 2015; 10(3): e0119332. doi: 10.1371/journal.pone.0119332.
- [6] Adejumo OA, Jinabhai C, Daniel O, et al. The effects of stigma and social support on the health-related quality of life of people with drug-resistant tuberculosis in Lagos, Nigeria. *Qual Life Res.* 2025; 34(5): 1305 – 1316. doi: 10.1007/s11136-024-03876-5.
- [7] Agbeko CK, Mallah MA, He B, et al. Mental health status and its impact on TB treatment and outcomes: a scoping review. *Front Public Health.* 2022; 10: 855515. doi: 10.3389/fpubh.2022.855515.

483 [8] World Health Organization. WHO operational handbook on tuberculosis: infection
 484 prevention and control. Geneva: WHO; 2022. Available from:
 485 [https://iris.who.int/server/api/core/bitstreams/292c8ced-94e3-4e19-9261-](https://iris.who.int/server/api/core/bitstreams/292c8ced-94e3-4e19-9261-2f8e00f4cea0/content)
 486 [2f8e00f4cea0/content](https://iris.who.int/server/api/core/bitstreams/292c8ced-94e3-4e19-9261-2f8e00f4cea0/content). Accessed 8 May 2026.

487 [9] National Health Commission of the People's Republic of China. Measures for the
 488 Ethical Review of Biomedical Research Involving Humans. Beijing: National Health
 489 Commission; 2016. Available from:
 490 [https://www.nhc.gov.cn/wjw/c100221/202201/558653bd3d1e4a078904152a30605](https://www.nhc.gov.cn/wjw/c100221/202201/558653bd3d1e4a078904152a30605a03.shtml)
 491 [a03.shtml](https://www.nhc.gov.cn/wjw/c100221/202201/558653bd3d1e4a078904152a30605a03.shtml). Accessed 15 May 2026.

492 [10] Romanowski K, Balshaw RF, Benedetti A, et al. Predicting tuberculosis relapse in
 493 patients treated with the standard 6-month regimen: an individual patient data meta-
 494 analysis. *Thorax*. 2019; 74(3): 291-297. doi: 10.1136/thoraxjnl-2018-212420.

495 [11] Afshar B, Carless J, Roche A, et al. Surveillance of tuberculosis (TB) cases
 496 attributable to relapse or reinfection in London, 2002-2015. *PLoS One*. 2019 Feb 15;
 497 14(2): e0211972. doi: 10.1371/journal.pone.0211972.

498 [12] Liu K, Chen S, Zhang Y, et al. Tuberculosis burden caused by migrant population in
 499 Eastern China: evidence from notification records in Zhejiang Province during 2013–
 500 2017. *BMC Infect Dis*. 2022; 22(1): 109. doi:10.1186/s12879-022-07071-5.

501 [13] Ki MS, Jeong D, Kang HY, et al. Real-world impact of the fixed-dose combination
 502 on improving treatment outcomes of drug-susceptible tuberculosis: a comparative
 503 study using multiyear national tuberculosis patient data. *BMJ Open Respir Res*. 2023;
 504 10(1): e001758. doi:10.1136/bmjresp-2023-001758.