healthcare

MDPI

Article

Machine Learning Prediction of Unsuccessful Tuberculosis Treatment Outcomes Using National Registry Data in Malaysia, 2014–2021

Dzulfitree Bin Ahmad [1], Shamsul Azhar Shah [1,*], Mohd Ihsani Bin Mahmood [2], Rabiah Abdul Kadir [3] and Muhammad Sufi Mohd Pauze [1]

[1] Department of Public Health Medicine, Faculty of Medicine, Universiti Kebangsaan Malaysia, Jalan Yaacob Latif, Bandar Tun Razak, Cheras, Kuala Lumpur 56000, Malaysia; dzulfitree@gmail.com (D.B.A.); p166271@siswa.ukm.edu.my (M.S.M.P.)
[2] Disease Control Division, Ministry of Health Malaysia, Putrajaya 62590, Malaysia; dr.ihsani@moh.gov.my
[3] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia; rabiahivi@ukm.edu.my
* Correspondence: drsham@hctm.ukm.edu.my; Tel.: +60-389215555

Highlights

What are the main findings?

- Among 193,862 tuberculosis treatment episodes registered in 2014–2021, 20.2% had an unsuccessful outcome; in multivariable logistic regression, the strongest adjusted associations were TB meningitis, having no DOT supervisor and HIV infection.
- Gradient boosting, XGBoost, LightGBM and a stacked ensemble had near-identical discrimination (ROC-AUC 0.752–0.754); SHAP was used to explain the model's predictions, not to identify causal risk factors.
- On a 2020–2021 temporal test set, gradient boosting improved ROC-AUC over logistic regression by about 0.01 (0.753 vs. 0.741); registry predictors gave moderate discrimination.

What are the implications of the main findings?

- Combining multivariable regression with interpretable machine learning can help national tuberculosis programmes identify population groups requiring enhanced clinical, adherence, and social support.
- Improved registry completeness and independent temporal and external validation are required before applying the model to individual clinical decisions or automated resource allocation.

Check for updates

Academic Editors: Bogdan Oancea, Adrian Pană and Cătălina Liliana Andrei

Received: 21 September 2026
Revised: 8 November 2026
Accepted: 9 November 2026
Published: 15 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Abstract

Background/Objectives: Unsuccessful tuberculosis (TB) treatment outcomes impede disease control. This study identified associated factors and evaluated machine-learning models using Malaysian national registry data. **Methods**: A retrospective cohort study used Malaysian National Tuberculosis Information System records registered from 2014 to 2021 with an evaluated treatment outcome (n = 193,862); registrations from 2022–2023 were excluded because follow-up was incomplete. Missing predictor values were retained as a separate 'Not recorded' category. Multivariable logistic regression with 17 prespecified predictors quantified adjusted associations. Logistic regression, decision tree, k-nearest neighbours, random forest, gradient boosting, XGBoost, LightGBM and a stacked ensemble were developed on 2014–2019 registrations ((n = 149,322) using patient-grouped cross-validation tuned for the area under the precision–recall curve (PR-AUC) and evaluated on a 2020–2021 temporal test set ((n = 44,540) with unsuccessful outcome as the positive

class, using ROC-AUC, PR-AUC, calibration, decision curves and classification metrics at an exploratory threshold of 0.20. SHAP was used for explainability. **Results**: Unsuccessful outcomes occurred in 39,172 episodes (20.2%), of which 55.6% were deaths. The strongest adjusted associations were observed for TB meningitis (aOR = 5.15 (95% CI: 4.72–5.61)), having no DOT supervisor (aOR = 4.72 (95% CI: 4.19–5.30)) and HIV-positive status (aOR = 3.36 (95% CI: 2.97–3.81)). The higher odds among non-Malaysian patients (aOR = 2.15) depended on classifying transfers out as unsuccessful. On the test set, ROC-AUC was 0.741 (95% CI: 0.735–0.746) for conventional logistic regression and 0.753 (95% CI: 0.747–0.758) for gradient boosting, the primary model; XGBoost, LightGBM and the stacked ensemble were within 0.001 of gradient boosting. At the 0.20 threshold, gradient boosting had a sensitivity of 0.696, specificity of 0.675, positive predictive value of 0.365 and F1 score of 0.479. Calibration was good (slope 0.997 for logistic regression and 0.959 for gradient boosting). SHAP identified ethnicity/citizenship, HIV status, age, sex and chest radiography as the largest contributors to model predictions. **Conclusions**: Routinely collected national registry variables provided moderate discrimination for unsuccessful TB treatment outcomes, and complex machine-learning models added only about 0.01 in ROC-AUC over logistic regression. Combining multivariable regression with interpretable machine learning provides complementary epidemiological and predictive insights; external validation and fairness evaluation are required before individual-level implementation.

Keywords: tuberculosis; treatment outcomes; machine learning; predictive modelling; national registry; risk stratification; Malaysia

1. Introduction

Tuberculosis (TB) remains the leading cause of death from a single infectious agent worldwide. In 2024, an estimated 10.7 million people developed TB and 1.23 million died from the disease [1]. Despite the availability of effective treatment, the global TB incidence rate declined by only 12% between 2015 and 2024, substantially below the 50% reduction targeted for 2025 [1]. The World Health Organization (WHO) End TB Strategy calls for reductions of 90% in TB incidence and 95% in TB deaths by 2035 relative to 2015 [2]. Malaysia continues to experience a substantial TB burden: 26,781 cases were reported in 2023, corresponding to a notification rate of 80.2 per 100,000 population, together with 2623 recorded deaths [3]. Malaysia's National Strategic Plan to End TB 2021–2030 requires notification of TB diagnosed or treated in public and private healthcare facilities and prioritises early detection, effective treatment and systematic programme monitoring [4].

Successful completion of treatment is essential to achieve cure, reduce mortality, interrupt transmission and prevent the emergence of drug resistance. WHO surveillance guidance records treatment outcomes as distinct categories, including cured, treatment completed, treatment failed, died, lost to follow-up and not evaluated [5]. Cure and treatment completion constitute treatment success. For epidemiological analyses, death, treatment failure, loss to follow-up or treatment interruption and outcomes that are not evaluated are frequently grouped as unsuccessful outcomes. However, these components are not clinically or programmatically interchangeable and may have different predictors and corresponding interventions. National registry studies in Malaysia have consistently shown that approximately one-fifth of patients experience an unsuccessful outcome [6,7]. Older age, male sex, non-Malaysian nationality, lower educational attainment, smoking, diabetes, HIV infection, absence of a bacille Calmette–Guérin scar and indicators of advanced disease have been associated with poorer outcomes [6,7]. Although these findings support

population-level programme planning, conventional association analyses do not necessarily provide sufficiently accurate individual risk estimates for directing additional support.

Machine-learning methods offer a potential means of analysing complex relationships and interactions within large healthcare datasets. However, their reported performance in TB outcome prediction varies considerably across populations, predictor sets and outcome definitions. A Malaysian study using 1228 records from Penang in 2017 found that a hyperparameter-tuned XGBoost model provided a modest improvement in accuracy over logistic regression, but the analysis was limited to one state and one year [8]. In a Brazilian national registry study involving 243,726 patients, logistic regression, random forest and LightGBM produced similar areas under the receiver operating characteristic curve of approximately 0.71–0.72 for predicting loss to follow-up [9]. Conversely, a dynamic modelling study using longitudinal data from Moldova showed that predictive performance improved as follow-up laboratory information accumulated [10]. A prospective Brazilian study also developed a comparatively simple seven-predictor regression model with a c-statistic of 0.77 and reported calibration measures [11]. Collectively, these findings suggest that the availability, timing and clinical relevance of predictors may be at least as important as algorithmic complexity.

The methodological quality and clinical readiness of existing TB prediction models remain contested. A systematic review of 37 TB treatment-outcome models found substantial variation in study populations, outcome definitions and analytical methods; fewer than half reported calibration, only 16% underwent external validation and all were judged to have a high risk of bias in the analysis domain [12]. A more recent systematic review and meta-analysis of 34 artificial-intelligence and machine-learning studies reported a pooled area under the curve of 0.836, but heterogeneity was extremely high, only eight studies performed external validation and only one was considered to have a low overall risk of bias [13]. More broadly, a systematic review of clinical prediction research found no consistent performance advantage of machine learning over logistic regression in comparisons at low risk of bias [14]. Consequently, good discrimination alone does not establish clinical usefulness. Transparent reporting of outcome definitions, missing-data management, model development, calibration, internal and external evaluation, subgroup performance and potential algorithmic bias is essential before a model can support healthcare decisions [15].

Evidence therefore remains limited on whether applying increasingly complex machine-learning algorithms to routinely collected Malaysian TB surveillance data provides meaningful predictive value beyond conventional regression. Previous national Malaysian studies primarily used conventional statistical modelling [6,7], while the available Malaysian machine-learning study was restricted to a single state and year [8]. The present study used National Tuberculosis Information System records collected throughout Malaysia for patients registered from 2014 to 2021. We aimed to identify factors associated with unsuccessful treatment outcomes and to evaluate whether logistic regression, decision tree, k-nearest neighbours, random forest, gradient boosting, XGBoost, LightGBM and a stacked ensemble provide predictive value beyond conventional multivariable logistic regression evaluated on the same temporal test set and to examine model contributions using explainability methods.

2. Materials and Methods

2.1. Study Design and Reporting

This retrospective national registry-based study used routinely collected secondary data from the Malaysian National Tuberculosis Information System (TBIS). The source extract covered records from 1 January 2014 to 31 March 2023; the analysis was restricted to registrations from 2014 to 2021 because most patients registered in 2022–2023 had not

completed treatment at the time of extraction (Section 2.2). The observational component was reported with reference to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, while the prediction-model component was reported in accordance with TRIPOD+AI guidance [15,16].

2.2. Data Source, Setting, and Study Population

Tuberculosis is a notifiable disease in Malaysia, and diagnosed cases are recorded in the Malaysian National Tuberculosis Information System (TBIS), a national surveillance registry maintained by the Disease Control Division, Ministry of Health Malaysia [4]. TBIS contains sociodemographic, socioeconomic, clinical, treatment-supervision and final treatment-outcome information collected from all Malaysian states and federal territories.

All TBIS records within the study period were assessed for eligibility. No sampling was performed because the study included the entire population of eligible registry records. The unit of analysis was an eligible TB registry record.

The source dataset contained 232,370 records and 37 variables. We excluded 33,139 records registered in 2022–2023 because follow-up was incomplete at extraction: 35.1% of 2022 registrations were still on treatment, and 27.6% (2022) and 98.3% (2023) of evaluated outcomes were unsuccessful, compared with 18.4–21.4% in 2014–2021. Of the remaining 199,231 records, we excluded 520 records for patients still on treatment, 4710 records for which the diagnosis had subsequently been changed, and 139 records with no recorded outcome. The final analytical cohort comprised 193,862 treatment episodes. No records were excluded because of missing predictor values; missing values were retained as a separate 'Not recorded' category (Section 2.4). The record-selection process is presented in Figure 1.

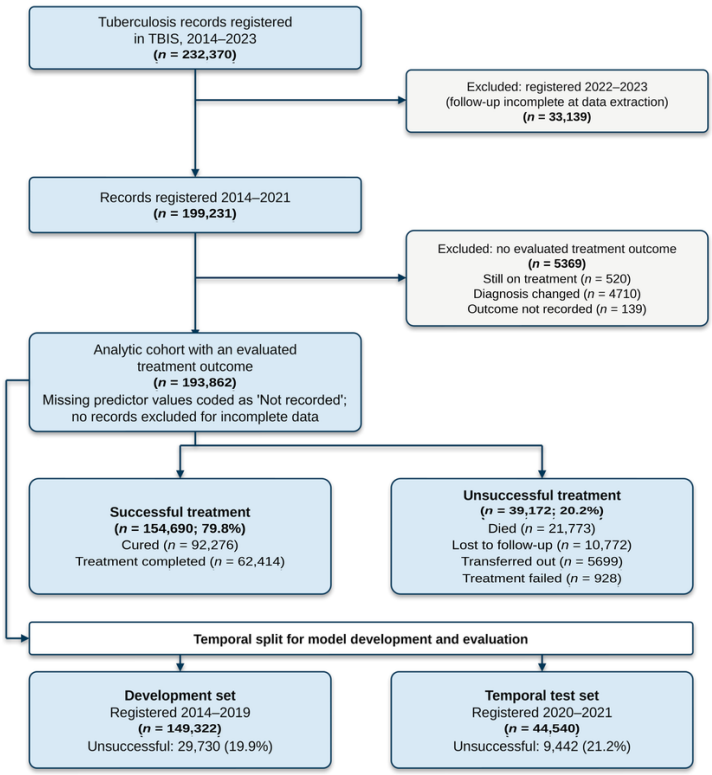


Figure 1. Selection of tuberculosis records for the analytical cohort and the temporal split used for prediction modelling. TBIS, Malaysian National Tuberculosis Information System. Percentages for treatment outcomes are based on the analytical cohort (= 193,862).

2.3. Outcome Definition

The study outcome was the final tuberculosis treatment outcome recorded in TBIS. The original registry outcome categories were consolidated into a binary endpoint. Successful treatment comprised cure or treatment completion. Unsuccessful treatment comprised death from any cause during treatment, treatment interruption or loss to follow-up, treatment failure, and transfer out without a documented successful outcome in the analysed registry extract. These classifications were based on the national registry coding and World Health Organization tuberculosis surveillance terminology [5].

Transferred-out records without a documented successful outcome were included in the study-specific unsuccessful composite because treatment success could not be confirmed from the available data. Under WHO definitions, such patients are classified as 'not evaluated' and are not counted as treatment success [5], and previous Malaysian national registry studies likewise counted transfers out or not-evaluated outcomes as unfavourable [6,7]; the classification also reflects the programme view that a transfer without a confirmed outcome is a break in continuity of care that warrants follow-up. Because some of these patients may have completed treatment elsewhere, all analyses were repeated excluding them (Section 2.7). Patients who were still receiving treatment, whose diagnosis had been changed, or whose outcome was not recorded were excluded because their final treatment outcomes were unavailable. The target event for prediction was unsuccessful treatment. The individual components were reported separately because they represent different clinical and programme processes; component-specific models were fitted as sensitivity analyses (Section 2.7), and no time-to-event model was fitted.

2.4. Candidate Predictors and Data Preparation

Eighteen candidate registry fields were selected based on their availability in TBIS, clinical and programme relevance, and evidence from previous tuberculosis treatment-outcome studies [6–13]. The candidate predictors comprised sociodemographic characteristics (age, sex, ethnicity, nationality and educational attainment); socioeconomic and contextual characteristics (employment status, healthcare-worker status, residential locality, and number of dependants); lifestyle and medical characteristics (smoking status, diabetes, HIV status, and bacille Calmette–Guérin scar status); and tuberculosis-related characteristics (tuberculous meningitis, miliary tuberculosis, tuberculosis case category, chest-radiograph status and the principal directly observed treatment (DOT) supervisor).

Predictor values were derived from the corresponding fields recorded in TBIS. Age was classified into seven predefined age groups. Free-text occupation entries were classified as employed or unemployed; homemakers, students or children, retired persons and prisoners or detainees, who were not in paid employment, were classified as unemployed. Because ethnicity was blank for almost all non-citizens (30,582 of 30,662 non-citizen records), ethnicity and citizenship were combined into a single variable (Malay, Chinese, Indian, other and non-Malaysian), in which the other category comprised Bumiputera Sabah, Bumiputera Sarawak, Orang Asli and other Malaysian ethnic groups, giving 17 analytic predictors. Number of dependants was grouped as 0–1, 2–4 and 5 or more, and chest radiograph findings at diagnosis as minimal or no lesion, moderately advanced, far advanced or not done. Missing predictor values were retained as a separate 'Not recorded' category, which for HIV status also included patients recorded as not tested; 'Not done' for chest radiography and 'None' for DOT supervision were recorded registry values and were analysed as distinct categories rather than as missing data. Residential locality was categorised as urban or rural. BCG-scar status was treated as a physical examination finding and was not assumed to be equivalent to documented vaccination history. No additional information was collected directly from patients.

Two prediction time points were defined. Fifteen predictors are recorded at registration or diagnosis, whereas HIV status and the DOT supervisor are recorded at or shortly after treatment initiation; no end-of-treatment information was used, but exact recording timestamps were not available in the extract. The main (early-treatment) models used all 17 predictors and are intended for use once HIV status and the intensive-phase DOT supervisor have been recorded, typically within the first weeks of treatment. Because a DOT supervisor recorded as 'None' or an HIV status that was not recorded (including not tested) may partly reflect early disengagement, these models may include information that is not available at treatment initiation. A baseline model using only the 15 predictors available at registration was therefore also fitted and reported in full (Section 2.7). The coding of the study outcome and, for each of the 17 analytic predictors derived from the 18 TBIS source fields, the operational definition, analytic categories, regression reference category and intended availability relative to the prediction time point are presented in Supplementary Table S1.

2.5. Conventional Statistical Analysis

Categorical variables were summarised using frequencies and percentages and compared using chi-square tests. Unsuccessful treatment was specified as the event of interest. Crude odds ratios were estimated using univariable logistic regression for information only. All 17 prespecified predictors were entered simultaneously into a multivariable logistic-regression model, without significance-based screening or data-driven variable selection. Adjusted odds ratios with Wald 95% confidence intervals and p values were reported. Statistical significance was defined as a two-sided p value < 0.05. The reference category for each predictor was prespecified and is detailed in Supplementary Table S1.The reference category used for each predictor is shown in Supplementary Table S1. As sensitivity analyses for missing data, the model was refitted using multiple imputation by chained equations (five imputed datasets) and complete-case analysis. Statistical analyses were conducted in Python 3.13.

These analyses were used to estimate adjusted associations rather than causal effects. The conventional (unpenalised) logistic-regression model was also fitted on the development data and evaluated on the same temporal test set as the machine-learning models, allowing a paired comparison of predictive performance.

2.6. Machine-Learning Model Development

Machine-learning analyses were conducted in Python 3.13 using scikit-learn 1.9.1, XGBoost 3.4.1 and LightGBM 4.7.0. The analytical cohort ($N = 193{,}862$) was divided chronologically into a development set of registrations from 2014–2019 ($n = 149{,}322$; 29,730 unsuccessful, 19.9%) and a temporal test set of registrations from 2020–2021 (($n = 44{,}540$; 9442 unsuccessful, 21.2%). No patient identifier appeared in both sets, and all cross-validation was grouped by patient identifier. Unsuccessful treatment outcome was designated as the positive class (coded 1), whereas successful treatment outcome was designated as the negative class.

Categorical predictors were one-hot encoded within scikit-learn pipelines that were fitted on the training folds only; no scaling was required because all predictors were categorical. The coding applied to the outcome and individual predictors is presented in Supplementary Table S1.

All 17 predictors described in Section 2.4 were used in every model; no data-driven feature selection was performed.

Eight individual models were evaluated: conventional (unpenalised) logistic regression, L2-penalised logistic regression, decision tree, k-nearest neighbours, random for-

est, gradient boosting (histogram-based), extreme gradient boosting (XGBoost) and light gradient-boosting machine (LightGBM). Hyperparameters were tuned by grid search in five-fold cross-validation stratified by outcome and grouped by patient in the development data, using PR-AUC as the selection criterion. The model with the highest cross-validated PR-AUC (gradient boosting, 0.465; LightGBM, 0.464; XGBoost, 0.463) was prespecified as the primary model before the test set was used. A stacked ensemble combined L2-penalised logistic regression, random forest, gradient boosting, XGBoost and LightGBM through a logistic-regression meta-learner fitted on out-of-fold predicted probabilities from the same grouped folds. An exploratory decision threshold of 0.20 was prespecified. It corresponds to an assumed cost ratio of 4:1 (threshold = $1/(1+4)$), reflecting the judgement that missing a patient who will have an unsuccessful outcome is more harmful than offering low-cost additional support, such as adherence counselling, closer follow-up or social-support referral, to a patient who would have succeeded. This ratio was not derived from empirical cost data and has not been confirmed with programme stakeholders. The threshold is close to the development-set outcome prevalence (19.9%), a common reference threshold when outcomes are imbalanced that gives a more even balance of sensitivity and specificity for models not corrected for class imbalance [17,18], and on out-of-fold development predictions it was also the threshold that minimised the assumed misclassification cost for gradient boosting and L2-penalised logistic regression. Threshold-independent measures were therefore treated as the primary performance results, and performance was also reported at alternative thresholds of 0.10, 0.15, 0.25, 0.30, 0.40 and 0.50 (implied cost ratios from 9:1 to 1:1). The test set was used once, for final evaluation. Final model performance was evaluated using the temporal test set.

2.7. Model Performance and Explainability

Performance on the temporal test set was assessed using ROC-AUC, PR-AUC, the Brier score and calibration (calibration plot, calibration intercept and slope, and observed-to-expected ratio), and, at the 0.20 threshold, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, balanced accuracy and accuracy, with confusion matrices. Confidence intervals (95%) were obtained from 500 stratified bootstrap resamples of the test set, and ROC-AUCs were compared using paired DeLong tests. Clinical usefulness was assessed using decision-curve analysis. Stability was assessed using three repeats of five-fold grouped cross-validation in the development data.

Unsuccessful treatment outcome was designated as the positive class for all metrics. Performance was also assessed by sex, age group, ethnicity/citizenship, citizenship, educational attainment and residential locality, reporting the number of patients and events, ROC-AUC, sensitivity, specificity and PPV at the 0.20 threshold, and calibration (O/E ratio and calibration slope), with bootstrap 95% confidence intervals (Poisson intervals for the O/E ratio). Between-category differences in sensitivity and false-positive rate at 0.20, with bootstrap 95% CIs, and the threshold each category would need to reach the overall sensitivity were also reported (Supplementary Table S8). For the component models, the intercept was updated on the 2020 test episodes, and the updated models were evaluated on the 2021 episodes (Supplementary Table S9). Sensitivity analyses (i) excluded transferred-out patients; (ii) modelled death and (iii) loss to follow-up, failure or transfer separately against successful treatment; (iv) removed HIV status and DOT supervisor, which are recorded at or after treatment initiation; (v) used complete cases only; (vi) removed ethnicity/citizenship and educational attainment; and (vii) evaluated each test year separately. For analyses (ii) and (iii), event counts, ROC-AUC, PR-AUC, the Brier score, calibration and classification metrics (at 0.20 and at the component's development-set prevalence) were reported; for analysis (iv), the resulting 15-predictor baseline model was reported in full.

To assess case-mix change and data drift, the development and temporal test cohorts were compared across all predictors, including 'Not recorded' categories, using standardised differences, with an absolute value above 0.10 regarded as meaningful.

SHapley Additive exPlanations (SHAP) were computed for the primary gradient-boosting model using TreeSHAP, and for L2-penalised logistic regression using exact linear SHAP, on 3000 randomly selected test episodes [19,20]; values for the one-hot columns of each predictor were summed. Positive SHAP values indicated contributions towards an unsuccessful-treatment prediction, whereas negative SHAP values indicated contributions towards a successful-treatment prediction. Global importance was summarised using mean absolute SHAP values, and permutation importance (decrease in test PR-AUC) was used as a model-agnostic check. SHAP findings were interpreted as explanations of model predictions rather than evidence of causal effects.

2.8. Ethical Approval and Data Governance

Ethical approval was obtained from the Research Ethics Committee of Universiti Kebangsaan Malaysia (project code JEP-2022-493; reference UKM PPI/111/8/JEP-2022-493; approved on 28 July 2022) and the Medical Research and Ethics Committee, Ministry of Health Malaysia (NMRR ID-22-02644-XU1 [IIR]; approved on 22 February 2023). Permission to access and use data from the Malaysian National Tuberculosis Information System was obtained from the Disease Control Division, Ministry of Health Malaysia, before commencement of the study. The study was conducted in accordance with the Declaration of Helsinki and applicable Malaysian ethical requirements.

This study involved the secondary analysis of de-identified registry data, with no direct contact with participants. The requirement for individual informed consent was waived by the approving ethics committees. The analytical data were stored on password-protected systems and were accessible only to authorised members of the research team. In accordance with the approved research protocol, the data will be retained for seven years following completion of the study and securely destroyed thereafter. Study findings are reported exclusively in aggregate form, and no identifiable participant information is presented.

2.9. Availability of Data, Code and Protocol

The individual-level data analysed in this study were obtained from the Malaysian National Tuberculosis Information System and remain confidential under the governance of the Disease Control Division, Ministry of Health Malaysia. The authors are not authorised to deposit the data in a public repository or redistribute them to third parties. Requests for data access should be directed to the Disease Control Division and remain subject to approval by the data custodian and compliance with applicable governmental, institutional, ethical and data-governance requirements. No public data accession number is available.

2.10. Use of Generative Artificial Intelligence

During the preparation of this manuscript, the authors used OpenAI Codex (accessed September 2026) to assist with language refinement, consistency checks against author-provided source documents, and the editorial preparation of tables and figures from established study data and outputs. The tool was not used to generate the original study data or to perform the original statistical or machine-learning analyses. All AI-assisted content was reviewed, edited and verified by the authors, who take full responsibility for the content of the manuscript.

3. Results

The results are presented according to cohort selection and treatment-outcome distribution, characteristics of the analytical cohort, factors associated with unsuccessful treatment outcomes, machine-learning model performance, calibration, clinical utility and sensitivity analyses, and SHAP-based explainability of the primary model.

3.1. Cohort Selection and Treatment Outcomes

A total of 232,370 tuberculosis records registered in the Malaysian National Tuberculosis Information System between 2014 and 2023 were assessed for eligibility. Of these, 33,139 records registered in 2022–2023 were excluded because follow-up was incomplete. Among the remaining 199,231 records, 520 were excluded because treatment was ongoing, 4710 because the tuberculosis diagnosis had subsequently been changed, and 139 because no outcome was recorded. The final analytical cohort comprised 193,862 treatment episodes, representing 83.4% of the source dataset (Figure 1). The development set comprised 149,322 episodes registered in 2014–2019 and the temporal test set 44,540 episodes registered in 2020–2021. Compared with the development cohort, the test cohort differed most in the recording of employment status, which was 'Not recorded' for 1.6% vs. 18.2% of episodes (standardised difference −0.58), with a corresponding increase in unemployed patients (54.1% vs. 42.1%; 0.24), and in the proportion with zero or one dependant (44.0% vs. 36.1%; 0.16). The test cohort also had a higher proportion of unsuccessful outcomes (21.2% vs. 19.9%), of which 62.2% vs. 53.5% were deaths (Supplementary Table S2).

Among the 193,862 episodes included in the analysis, 154,690 (79.8%) had a successful treatment outcome, and 39,172 (20.2%) had an unsuccessful treatment outcome. Successful outcomes comprised 92,276 patients classified as cured, and 62,414 classified as having completed treatment. Unsuccessful outcomes comprised 21,773 deaths (55.6% of unsuccessful outcomes), 10,772 patients lost to follow-up (27.5%), 928 treatment failures (2.4%) and 5699 patients transferred out without a documented successful treatment outcome (14.5%).

3.2. Characteristics of the Analytical Cohort

Among the 193,862 episodes included in the analytical cohort, 124,627 (64.3%) were male, 88,988 (45.9%) were aged 45 years or older, 86,154 (44.4%) were of Malay ethnicity, 25,453 (13.1%) were non-Malaysian, 107,318 (55.4%) resided in urban areas, and 79,007 (40.8%) were employed. Smoking was reported in 61,294 episodes (31.6%), HIV-positive status in 1208 (0.6%), diabetes mellitus in 36,334 (18.7%) and absence of a BCG scar in 33,184 (17.1%). The complete distribution of patient characteristics according to treatment outcome is presented in Table 1.

Table 1. Characteristics of the analytical cohort according to treatment outcome (N = 193,862).

Characteristic	Overall, *n* (%)	Successful, *n* (%) (*N* = 154,690)	Unsuccessful, *n* (%) (*N* = 39,172)	*p* Value
Age, years				<0.001
0–14	5206 (2.7)	4845 (3.1)	361 (0.9)	
15–24	29,155 (15.0)	24,929 (16.1)	4226 (10.8)	
25–34	36,893 (19.0)	29,694 (19.2)	7199 (18.4)	
35–44	32,945 (17.0)	25,514 (16.5)	7431 (19.0)	
45–54	32,508 (16.8)	26,141 (16.9)	6367 (16.3)	
55–64	29,834 (15.4)	23,826 (15.4)	6008 (15.3)	
≥65	26,646 (13.7)	19,172 (12.4)	7474 (19.1)	

Table 1. *Cont.*

Characteristic	Overall, *n* (%)	Successful, *n* (%) (*N* = 154,690)	Unsuccessful, *n* (%) (*N* = 39,172)	*p* Value
Not recorded	675 (0.3)	569 (0.4)	106 (0.3)	
Sex				<0.001
Male	124,627 (64.3)	95,616 (61.8)	29,011 (74.1)	
Female	69,235 (35.7)	59,074 (38.2)	10,161 (25.9)	
Ethnicity/citizenship				<0.001
Malay	86,154 (44.4)	69,536 (45.0)	16,618 (42.4)	
Chinese	28,225 (14.6)	23,264 (15.0)	4961 (12.7)	
Indian	10,947 (5.6)	8081 (5.2)	2866 (7.3)	
Other	43,043 (22.2)	37,296 (24.1)	5747 (14.7)	
Non-Malaysian	25,453 (13.1)	16,500 (10.7)	8953 (22.9)	
Not recorded	40 (0.0)	13 (0.0)	27 (0.1)	
Educational attainment				<0.001
None or primary	67,708 (34.9)	50,387 (32.6)	17,321 (44.2)	
Secondary	96,460 (49.8)	78,252 (50.6)	18,208 (46.5)	
Tertiary	27,686 (14.3)	24,794 (16.0)	2892 (7.4)	
Other	2008 (1.0)	1257 (0.8)	751 (1.9)	
Employment status				<0.001
Employed	79,007 (40.8)	64,983 (42.0)	14,024 (35.8)	
Unemployed	87,020 (44.9)	67,709 (43.8)	19,311 (49.3)	
Not recorded	27,835 (14.4)	21,998 (14.2)	5837 (14.9)	
Healthcare worker				<0.001
No	190,877 (98.5)	151,841 (98.2)	39,036 (99.7)	
Yes	2985 (1.5)	2849 (1.8)	136 (0.3)	
Residential locality				<0.001
Urban	107,318 (55.4)	84,286 (54.5)	23,032 (58.8)	
Rural	85,068 (43.9)	69,275 (44.8)	15,793 (40.3)	
Not recorded	1476 (0.8)	1129 (0.7)	347 (0.9)	
Smoking				<0.001
No	132,568 (68.4)	108,163 (69.9)	24,405 (62.3)	
Yes	61,294 (31.6)	46,527 (30.1)	14,767 (37.7)	
Diabetes mellitus				<0.001
No	157,528 (81.3)	126,276 (81.6)	31,252 (79.8)	
Yes	36,334 (18.7)	28,414 (18.4)	7920 (20.2)	
HIV status				<0.001
Negative	172,364 (88.9)	141,734 (91.6)	30,630 (78.2)	
Positive	1208 (0.6)	678 (0.4)	530 (1.4)	
Not recorded	20,290 (10.5)	12,278 (7.9)	8012 (20.5)	
BCG scar				<0.001
Present	160,678 (82.9)	131,777 (85.2)	28,901 (73.8)	
Absent	33,184 (17.1)	22,913 (14.8)	10,271 (26.2)	
Miliary tuberculosis				<0.001
No	192,206 (99.1)	153,723 (99.4)	38,483 (98.2)	
Yes	1656 (0.9)	967 (0.6)	689 (1.8)	
Tuberculous meningitis				<0.001
No	191,232 (98.6)	153,475 (99.2)	37,757 (96.4)	
Yes	2630 (1.4)	1215 (0.8)	1415 (3.6)	
Case category				<0.001
New	178,643 (92.1)	144,014 (93.1)	34,629 (88.4)	
Relapse	10,095 (5.2)	7672 (5.0)	2423 (6.2)	
After failure	876 (0.5)	580 (0.4)	296 (0.8)	
After loss to follow-up	4248 (2.2)	2424 (1.6)	1824 (4.7)	
Number of dependants				<0.001

Table 1. *Cont.*

Characteristic	Overall, *n* (%)	Successful, *n* (%) (*N* = 154,690)	Unsuccessful, *n* (%) (*N* = 39,172)	*p* Value
0–1	73,527 (37.9)	54,068 (35.0)	19,459 (49.7)	
2–4	68,451 (35.3)	57,457 (37.1)	10,994 (28.1)	
≥5	40,727 (21.0)	34,280 (22.2)	6447 (16.5)	
Not recorded	11,157 (5.8)	8885 (5.7)	2272 (5.8)	
Chest radiograph at diagnosis				<0.001
Minimal/no lesion	116,087 (59.9)	95,479 (61.7)	20,608 (52.6)	
Moderately advanced	65,155 (33.6)	50,310 (32.5)	14,845 (37.9)	
Far advanced	8338 (4.3)	5498 (3.6)	2840 (7.3)	
Not done	4282 (2.2)	3403 (2.2)	879 (2.2)	
DOT supervisor, intensive phase				<0.001
Health staff	122,437 (63.2)	95,028 (61.4)	27,409 (70.0)	
Family member	67,123 (34.6)	57,493 (37.2)	9630 (24.6)	
Other	2722 (1.4)	1647 (1.1)	1075 (2.7)	
None	1580 (0.8)	522 (0.3)	1058 (2.7)	

Values are n (%). Percentages are column percentages within the full cohort or corresponding outcome group. *p* values are from chi-square tests. 'Not recorded' denotes values missing in TBIS, which were retained as a separate category. BCG, Bacillus Calmette–Guérin; DOT, directly observed treatment; HIV, human immunodeficiency virus.

Compared with patients who achieved a successful treatment outcome, those with an unsuccessful outcome had higher proportions of patients aged 65 years or older (19.1% vs. 12.4%), males (74.1% vs. 61.8%), non-Malaysian patients (22.9% vs. 10.7%), patients with no or primary education (44.2% vs. 32.6%), unemployed patients (49.3% vs. 43.8%), smokers (37.7% vs. 30.1%), patients with HIV-positive status (1.4% vs. 0.4%), patients with diabetes mellitus (20.2% vs. 18.4%) and patients without a recorded BCG scar (26.2% vs. 14.8%). In the unadjusted comparisons, all characteristics presented in Table 1 differed significantly between the successful- and unsuccessful-treatment groups (all $p < 0.001$).

3.3. Factors Associated with Unsuccessful Treatment Outcomes

In the multivariable logistic-regression analysis including all 17 predictors (Table 2), compared with patients aged 25–34 years, the adjusted odds of an unsuccessful treatment outcome were lower among patients aged 0–14 years (aOR = 0.31 (95% CI: 0.28–0.35)) and 15–24 years (aOR = 0.81 (95% CI: 0.78–0.85)) and highest among those aged 65 years or older (aOR = 1.96 (95% CI: 1.87–2.05)). The strongest associations were observed for TB meningitis (aOR = 5.15 (95% CI: 4.72–5.61)) and having no DOT supervisor during the intensive phase compared with a health-staff supervisor (aOR = 4.72 (95% CI: 4.19–5.30)).

Table 2. Multivariable logistic-regression factors associated with unsuccessful tuberculosis treatment outcomes ((*N* = 193,862).

Variable and Category	Adjusted Odds Ratio	95% Confidence Interval	*p* Value
Age, years			
0–14	0.31	0.28–0.35	<0.001
15–24	0.81	0.78–0.85	<0.001
25–34 (reference)	1.00	—	—
35–44	1.11	1.07–1.16	<0.001
45–54	1.07	1.02–1.11	0.003
55–64	1.22	1.16–1.27	<0.001

Table 2. *Cont.*

Variable and Category	Adjusted Odds Ratio	95% Confidence Interval	*p* Value
≥65	1.96	1.87–2.05	<0.001
Not recorded	0.55	0.43–0.69	<0.001
Sex			
Male (reference)	1.00	—	—
Female	0.65	0.63–0.67	<0.001
Ethnicity/citizenship			
Malay (reference)	1.00	—	—
Chinese	0.66	0.64–0.69	<0.001
Indian	1.31	1.24–1.38	<0.001
Other	0.82	0.79–0.85	<0.001
Non-Malaysian	2.15	2.05–2.26	<0.001
Not recorded	6.69	3.30–13.58	<0.001
Educational attainment			
None or primary	1.16	1.13–1.20	<0.001
Secondary (reference)	1.00	—	—
Tertiary	0.62	0.59–0.65	<0.001
Other	1.84	1.66–2.03	<0.001
Employment status			
Employed (reference)	1.00	—	—
Unemployed	1.29	1.25–1.32	<0.001
Not recorded	1.17	1.13–1.21	<0.001
Healthcare worker			
No (reference)	1.00	—	—
Yes	0.43	0.36–0.51	<0.001
Residential locality			
Urban (reference)	1.00	—	—
Rural	0.79	0.77–0.81	<0.001
Not recorded	0.98	0.85–1.11	0.714
Smoking			
No (reference)	1.00	—	—
Yes	1.10	1.07–1.14	<0.001
Diabetes mellitus			
No (reference)	1.00	—	—
Yes	1.09	1.05–1.12	<0.001
HIV status			
Negative (reference)	1.00	—	—
Positive	3.36	2.97–3.81	<0.001
Not recorded	2.80	2.71–2.90	<0.001
BCG scar			
Present (reference)	1.00	—	—
Absent	1.19	1.15–1.24	<0.001
Miliary tuberculosis			
No (reference)	1.00	—	—
Yes	1.95	1.74–2.18	<0.001
Tuberculous meningitis			
No (reference)	1.00	—	—
Yes	5.15	4.72–5.61	<0.001
Case category			
New (reference)	1.00	—	—
Relapse	1.19	1.14–1.26	<0.001
After failure	1.85	1.59–2.16	<0.001
After loss to follow-up	2.71	2.53–2.90	<0.001
Number of dependants			
0–1 (reference)	1.00	—	—

Table 2. *Cont.*

Variable and Category	Adjusted Odds Ratio	95% Confidence Interval	*p* Value
2–4	0.66	0.64–0.68	<0.001
≥5	0.65	0.63–0.67	<0.001
Not recorded	0.84	0.80–0.89	<0.001
Chest radiograph at diagnosis			
Minimal/no lesion (reference)	1.00	—	—
Moderately advanced	1.28	1.25–1.31	<0.001
Far advanced	2.14	2.03–2.25	<0.001
Not done	1.19	1.10–1.30	<0.001
DOT supervisor, intensive phase			
Health staff (reference)	1.00	—	—
Family member	0.60	0.59–0.62	<0.001
Other	1.46	1.34–1.59	<0.001
None	4.72	4.19–5.30	<0.001

Adjusted odds ratios are from one model including all 17 predictors and are reported to two decimal places. The reference category is shown explicitly for each predictor. 'Not recorded' categories were included in the model. BCG, Bacillus Calmette–Guérin; DOT, directly observed treatment; HIV, human immunodeficiency virus.

Higher adjusted odds of an unsuccessful treatment outcome were also observed for HIV-positive status (aOR = 3.36 (95% CI: 2.97–3.81)) and an unrecorded HIV status, including not tested (aOR = 2.80 (95% CI: 2.71–2.90)), compared with HIV-negative status, retreatment after loss to follow-up (aOR = 2.71 (95% CI: 2.53–2.90)) or after treatment failure (aOR = 1.85 (95% CI: 1.59–2.16)) compared with new cases, non-Malaysian citizenship compared with Malay ethnicity (aOR = 2.15 (95% CI: 2.05–2.26)), far-advanced chest radiography findings compared with minimal or no lesion (aOR = 2.14 (95% CI: 2.03–2.25)), miliary tuberculosis (aOR = 1.95 (95% CI: 1.74–2.18)), Indian ethnicity (aOR = 1.31 (95% CI: 1.24–1.38)), unemployed patients compared with employed patients (aOR = 1.29 (95% CI: 1.25–1.32)), absence of a BCG scar (aOR = 1.19 (95% CI: 1.15–1.24)), no or primary education compared with secondary education (aOR = 1.16 (95% CI: 1.13–1.20)), smoking (aOR = 1.10 (95% CI: 1.07–1.14)) and diabetes mellitus (aOR = 1.09 (95% CI: 1.05–1.12)).

Lower adjusted odds of an unsuccessful treatment outcome were observed among healthcare workers (aOR = 0.43 (95% CI: 0.36–0.51)), patients whose DOT supervisor was a family member (aOR = 0.60 (95% CI: 0.59–0.62)), patients with tertiary education (aOR = 0.62 (95% CI: 0.59–0.65)), female patients (aOR = 0.65 (95% CI: 0.63–0.67)), patients with two to four dependants (aOR = 0.66 (95% CI: 0.64–0.68)), patients of Chinese ethnicity (aOR = 0.66 (95% CI: 0.64–0.69)), rural residents (aOR = 0.79 (95% CI: 0.77–0.81)) and patients of other Malaysian ethnicity (aOR = 0.82 (95% CI: 0.79–0.85)). An unrecorded residential locality was the only category not significantly associated with the outcome ($p = 0.714$).

3.4. Machine-Learning Model Performance

In the development data, gradient boosting had the highest cross-validated PR-AUC (0.4647), followed by LightGBM (0.4636), XGBoost (0.4628) and random forest (0.4485), and was therefore the primary model. In the stacked ensemble, the meta-learner weights (log-odds scale) were 2.60 for gradient boosting, 1.78 for random forest, 1.34 for LightGBM, 0.34 for XGBoost and 0.09 for L2-penalised logistic regression.

On the temporal test set ((*n* = 44,540; prevalence 0.212), ROC-AUC ranged from 0.729 (k-nearest neighbours) to 0.754 (XGBoost), and PR-AUC from 0.418 (k-nearest neighbours) to 0.470 (XGBoost and LightGBM), compared with a prevalence baseline of 0.212 (Table 3). At the conventional threshold of 0.5, accuracy was 0.795–0.803, close to the no-information accuracy of 0.788, but sensitivity was only 0.095–0.231; classification metrics are therefore reported at the prespecified exploratory 0.20 threshold (Table 4), at which 404 of every 1000 test patients were flagged by gradient boosting. Across alternative thresholds (Supplementary Table S3), raising the threshold to 0.30 reduced the sensitivity of gradient boosting to 0.470 and increased PPV to 0.453, whereas lowering it to 0.10 increased sensitivity to 0.909 and reduced PPV to 0.277.

Table 3. Discrimination and calibration of the models in the temporal test set (2020–2021; (*N* = 44,540).

Model	ROC-AUC(95% CI)	PR-AUC(95% CI)	Brier Score	Calibration Intercept	Calibration Slope	O/E Ratio
Logistic regression (conventional)	0.741 (0.735–0.746)	0.455 (0.446–0.465)	0.144	0.023	0.997	1.02
Logistic regression (L2)	0.740 (0.735–0.745)	0.454 (0.445–0.464)	0.144	0.025	0.996	1.02
Decision tree	0.732 (0.726–0.737)	0.438 (0.429–0.447)	0.147	0.025	0.710	1.02
K-nearest neighbours	0.729 (0.723–0.734)	0.418 (0.409–0.426)	0.149	0.228	0.953	1.18
Random forest	0.743 (0.737–0.748)	0.450 (0.440–0.459)	0.145	0.035	1.183	1.02
Gradient boosting (HistGB) a	0.753 (0.747–0.758)	0.469 (0.459–0.478)	0.142	0.028	0.959	1.02
XGBoost	0.754 (0.748–0.759)	0.470 (0.461–0.479)	0.142	0.030	0.975	1.02
LightGBM	0.753 (0.747–0.758)	0.470 (0.460–0.479)	0.142	0.030	0.973	1.02
Stacked ensemble	0.752 (0.747–0.758)	0.469 (0.459–0.478)	0.143	0.024	0.943	1.02

Values in parentheses are 95% confidence intervals from 500 stratified bootstrap resamples. All models were fitted on the 2014–2019 development set and evaluated on the same test episodes. Ideal calibration corresponds to an intercept of 0, a slope of 1 and an observed-to-expected (O/E) ratio of 1. A primary model, chosen by the highest cross-validated PR-AUC in the development data. PR-AUC, area under the precision–recall curve (prevalence baseline 0.212); ROC-AUC, area under the receiver operating characteristic curve; XGBoost, extreme gradient boosting; LightGBM, light gradient-boosting machine.

Table 4. Classification performance at the prespecified exploratory threshold of 0.20 in the temporal test set ((*N* = 44,540).

Model	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 Score	Balanced Accuracy	TP/FP/TN/FN
Logistic regression (conventional)	0.681 (0.673–0.690)	0.666 (0.661–0.671)	0.354 (0.350–0.359)	0.886 (0.883–0.889)	0.466	0.674	6432/11,721/ 23,377/3010
Logistic regression (L2)	0.680 (0.670–0.688)	0.667 (0.662–0.672)	0.354 (0.349–0.359)	0.886 (0.883–0.888)	0.466	0.673	6416/11,695/ 23,403/3026
Decision tree	0.688 (0.678–0.696)	0.650 (0.645–0.654)	0.346 (0.341–0.349)	0.885 (0.882–0.888)	0.460	0.669	6493/12,295/ 22,803/2949
K-nearest neighbours	0.624 (0.614–0.633)	0.708 (0.703–0.713)	0.365 (0.360–0.370)	0.875 (0.872–0.878)	0.461	0.666	5894/10,233/ 24,865/3548

Table 4. *Cont.*

Model	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 Score	Balanced Accuracy	TP/FP/TN/FN
Random forest	0.710 (0.701–0.719)	0.645 (0.640–0.650)	0.350 (0.345–0.354)	0.892 (0.889–0.895)	0.469	0.678	6707/12,470/ 22,628/2735
Gradient boosting (HistGB) a	0.696 (0.686–0.706)	0.675 (0.670–0.680)	0.365 (0.360–0.370)	0.892 (0.889–0.895)	0.479	0.685	6575/11,424/ 23,674/2867
XGBoost	0.688 (0.679–0.697)	0.682 (0.677–0.687)	0.368 (0.363–0.372)	0.890 (0.887–0.893)	0.479	0.685	6500/11,174/ 23,924/2942
LightGBM	0.698 (0.688–0.707)	0.671 (0.666–0.676)	0.363 (0.358–0.368)	0.892 (0.889–0.895)	0.478	0.684	6586/11,547/ 23,551/2856
Stacked ensemble	0.626 (0.616–0.635)	0.736 (0.732–0.741)	0.390 (0.384–0.395)	0.880 (0.877–0.882)	0.480	0.681	5907/9249/ 25,849/3535

Values in parentheses are 95% confidence intervals from 500 stratified bootstrap resamples. Unsuccessful treatment was designated as the positive class as a primary model. FN, false negatives; FP, false positives; NPV, negative predictive value; PPV, positive predictive value; TN, true negatives; TP, true positives.

Gradient boosting achieved a ROC-AUC of 0.753 (95% CI: 0.747–0.758) and a PR-AUC of 0.469 (95% CI: 0.459–0.478). At the 0.20 threshold, its sensitivity was 0.696 (95% CI: 0.686–0.706), specificity 0.675 (95% CI: 0.670–0.680), PPV 0.365 (95% CI: 0.360–0.370), NPV 0.892 (95% CI: 0.889–0.895), F1 score 0.479 and balanced accuracy 0.685 (true positives 6575, false positives 11,424, true negatives 23,674, false negatives 2867). XGBoost, LightGBM and the stacked ensemble differed from gradient boosting by 0.001 or less in ROC-AUC (stacked ensemble vs. gradient boosting, difference 0.0008; DeLong $p = 0.001$). The stacked ensemble had a specificity of 0.736 and a sensitivity of 0.626 at the same threshold.

The conventional logistic-regression model, evaluated on the same test episodes, had a ROC-AUC of 0.741 (95% CI: 0.735–0.746) and a PR-AUC of 0.455; gradient boosting exceeded it by 0.012 (DeLong $p < 0.001$). At the 0.20 threshold, conventional logistic regression had a sensitivity of 0.681, specificity of 0.666, PPV of 0.354 and NPV of 0.886.

3.5. Calibration, Clinical Utility, Subgroups and Sensitivity Analyses

Calibration in the test set was good for the logistic-regression and boosting models (Table 3). Conventional logistic regression had a calibration intercept of 0.023, slope of 0.997, O/E ratio of 1.02 and Brier score of 0.144; gradient boosting had an intercept of 0.028, slope of 0.959, O/E ratio of 1.02 and Brier score of 0.142. The decision tree (slope 0.710) gave predictions that were too extreme, the random forest (slope 1.183) predictions that were too conservative, and k-nearest neighbours underestimated risk (O/E ratio 1.18). In decision-curve analysis, the net benefit at the 0.20 threshold was 0.0835 for gradient boosting, 0.079 for conventional logistic regression and 0.015 for treating all patients; all main models had higher net benefit than treating all or no patients between thresholds of about 0.10 and 0.50. In three repeats of five-fold grouped cross-validation, ROC-AUC was 0.7415 (SD 0.0033) for conventional logistic regression and 0.7556 (SD 0.0032) for gradient boosting.

Discrimination of gradient boosting varied by subgroup (ROC-AUC from 0.656 for patients aged 65 years or older to 0.777 for those aged 25–34 years, among subgroups with at least 1000 patients; from 0.726 for Indian to 0.743 for non-Malaysian patients across ethnicity/citizenship groups). Predicted risk was overestimated for non-Malaysian patients (O/E ratio 0.90) and for children aged 0–14 years (O/E ratio 0.80). At the 0.20 threshold, sensitivity ranged from 0.353 (95% CI: 0.316–0.390; 705 events) among patients with tertiary education to 0.930 (95% CI: 0.920–0.941; 2196 events) among non-Malaysian patients (Supplementary Table S5). These differences largely followed the distribution of predicted risk: 79.3% of non-Malaysian patients and 13.5% of patients with tertiary education had a

predicted risk of 0.20 or more. The largest between-category differences were by ethnicity/citizenship (sensitivity 0.930 for non-Malaysian vs. 0.442 for other patients; false-positive rate 0.721 vs. 0.144) and educational attainment (sensitivity 0.805 for none/primary vs. 0.353 for tertiary education); to reach the overall sensitivity of 0.696, the threshold would have needed to be 0.093 for patients with tertiary education and 0.363 for non-Malaysian patients (Supplementary Table S8).

In sensitivity analyses, test ROC-AUC for conventional logistic regression and gradient boosting was 0.738 and 0.744 when transferred-out patients were excluded ((*n* = 188,163; 33,473 unsuccessful, 17.8%); 0.791 and 0.793 for death versus successful treatment; 0.783 and 0.795 for loss to follow-up, failure or transfer versus successful treatment; 0.717 and 0.730 for the 15-predictor baseline model without HIV status and DOT supervisor; 0.727 and 0.741 in complete cases (138,231 of 193,862 episodes, 71.3%); and 0.725 and 0.737 without ethnicity/citizenship and educational attainment. In the individual test years, ROC-AUC was 0.736 and 0.749 in 2020 and 0.746 and 0.758 in 2021. Adjusted odds ratios were similar with multiple imputation and complete-case analysis; for non-Malaysian citizenship, the aOR was 1.98 (95% CI: 1.89–2.08) and 2.13 (95% CI: 2.01–2.25), respectively. However, the aOR for non-Malaysian citizenship fell to 0.88 (95% CI: 0.83–0.93) when transferred-out patients were excluded (full model in Supplementary Table S6) and was 1.22 (95% CI: 1.14–1.31) in the death-only model. For death versus successful treatment (5876 test events; prevalence 0.143), gradient boosting had a PR-AUC of 0.408 and a calibration slope of 0.922; for loss to follow-up, failure or transfer (3566 test events; prevalence 0.092), these were 0.346 and 0.936. Calibration-in-the-large differed between the component models (O/E ratio 1.16 for death and 0.86 for the other outcomes), consistent with the larger share of deaths among unsuccessful outcomes in the test period (Supplementary Tables S2 and S4). Updating the intercept on 2020 episodes improved the O/E ratio in 2021 from 1.20 to 1.07 for death and from 0.82 to 0.92 for the other outcomes, without changing discrimination (Supplementary Table S9). The baseline model had a PR-AUC of 0.399 for logistic regression and 0.420 for gradient boosting, with calibration slopes of 1.000 and 0.973 (Supplementary Table S7).

3.6. Explainability of the Primary Model

SHAP analysis examined the contribution of the 17 predictors to the predictions of the primary gradient-boosting model (Figure 2). Mean absolute SHAP values were largest for ethnicity/citizenship (0.261), HIV status (0.228), age group (0.193), sex (0.188) and chest radiograph (0.178). Permutation importance identified HIV status (0.063 decrease in test PR-AUC), ethnicity/citizenship (0.059) and DOT supervisor (0.049) as the most influential predictors.

Positive SHAP values shifted the model output towards an unsuccessful treatment outcome, whereas negative SHAP values shifted the output towards a successful treatment outcome. Among the 20 predictor categories with the largest contributions (Figure 2b), non-Malaysian citizenship, zero or one dependant, age 65 years or older, far-advanced chest radiography, absence of a BCG scar, no or primary education, unemployment and urban residence generally shifted predictions towards an unsuccessful treatment outcome. Negative HIV status, female sex, a family member as DOT supervisor, minimal or no lesion on chest radiography, tertiary education, Chinese or other ethnicity, employment, age below 25 years, non-smoking and new-case status generally shifted predictions towards a successful outcome.

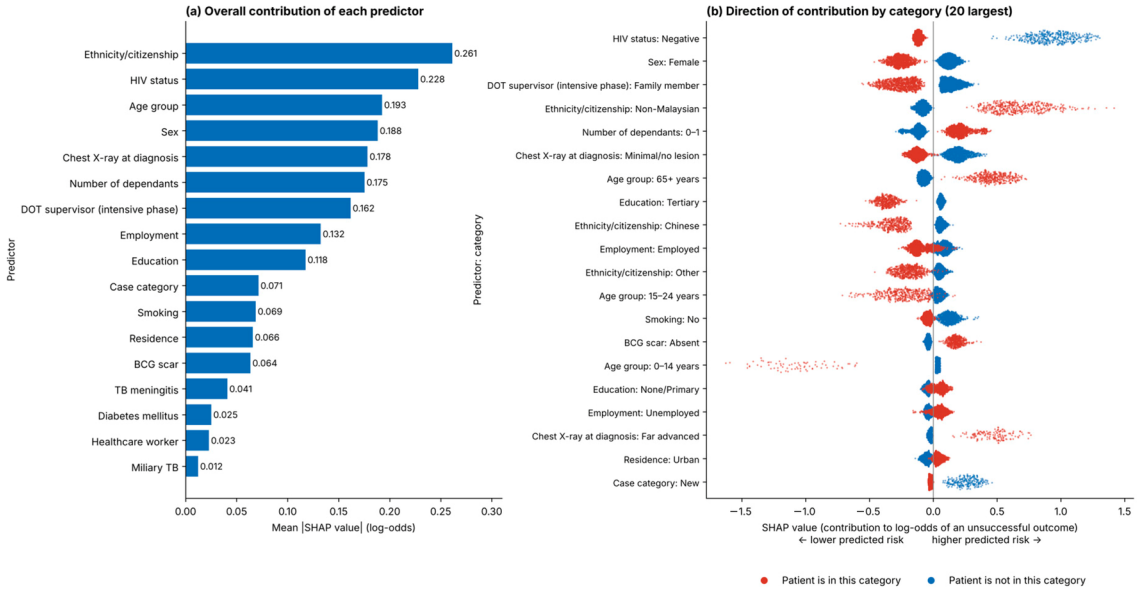


Figure 2. SHAP explanation of the primary gradient-boosting model in 3000 test episodes. (**a**) Mean absolute SHAP value for each predictor (log-odds scale). (**b**) SHAP values for the 20 predictor categories with the largest mean absolute contribution; each point is one patient, shown in red when the patient is in that category and in blue otherwise. Values above zero increase, and values below zero decrease, the predicted risk of an unsuccessful outcome. SHAP values explain the model's predictions and do not indicate causality. DOT, directly observed treatment; SHAP, SHapley Additive exPlanations.

These directions were broadly consistent with the adjusted odds ratios. These findings describe the behaviour of the fitted gradient-boosting model and should not be interpreted as evidence of causal effects or adjusted statistical associations.

4. Discussion

4.1. Principal Findings

In this national registry study of 193,862 tuberculosis treatment episodes registered in Malaysia in 2014–2021, 20.2% had an unsuccessful treatment outcome, and 55.6% of unsuccessful outcomes were deaths. Multivariable analysis identified a combination of demographic, socioeconomic, behavioural, and clinical factors associated with treatment outcome. The strongest associations with unsuccessful treatment were observed for TB meningitis (aOR = 5.15), having no DOT supervisor (aOR = 4.72), HIV-positive status (aOR = 3.36), an unrecorded HIV status (aOR = 2.80), retreatment after loss to follow-up (aOR = 2.71) and non-Malaysian citizenship (aOR = 2.15). Age 65 years or older, male sex, unemployment, urban residence, smoking, diabetes mellitus and absence of a BCG scar were also associated with higher adjusted odds of an unsuccessful outcome. Conversely, healthcare-worker status, tertiary education, Chinese ethnicity and having a family member as DOT supervisor were associated with lower adjusted odds.

Gradient boosting, the prespecified primary model, achieved a test-set ROC-AUC of 0.753 and PR-AUC of 0.469; at the 0.20 threshold, it identified 69.6% of unsuccessful outcomes, but only 36.5% of flagged patients had an unsuccessful outcome. XGBoost, LightGBM and the stacked ensemble performed almost identically (ROC-AUC 0.752–0.754), and conventional logistic regression had a ROC-AUC only 0.012 lower. The findings demonstrate that routinely collected national registry data contain a moderate predictive signal for unsuccessful tuberculosis treatment outcomes, and that more complex algorithms added little to it.

The SHAP analysis complemented the multivariable findings by showing how individual predictors contributed to the outputs of the gradient-boosting model. Ethnicity/citizenship, HIV status, age and sex had the largest overall contributions, while the directional SHAP patterns for age, sex, employment, education, HIV status, chest radiography, BCG-scar status and DOT supervisor were broadly consistent with the adjusted associations. This convergence supports a hybrid analytical framework in which multivariable logistic regression quantifies independent associations, machine learning characterises multivariable predictive patterns, and SHAP improves model interpretability. Collectively, the findings may help identify population groups that could benefit from enhanced monitoring, integrated clinical management, and additional social or programme support within Malaysia's national tuberculosis control strategy, subject to further validation.

4.2. Interpretation of Sociodemographic and Clinical Factors

The sociodemographic and clinical associations observed in this study were broadly consistent with previous Malaysian national registry studies, which identified older age, male sex, non-Malaysian citizenship, lower educational attainment, urban residence, smoking and HIV infection as factors associated with unsuccessful tuberculosis treatment outcomes [6,7]. The present study extends this evidence using ten years of national TBIS data and demonstrates that these associations remain important after simultaneous adjustment for demographic, socioeconomic, behavioural and clinical characteristics.

Older age remained associated with unsuccessful treatment after adjustment. Patients aged 65 years or older had about twice the adjusted odds of an unsuccessful outcome compared with those aged 25–34 years (aOR = 1.96), and the association was stronger for death alone (aOR = 4.34 in the death-only model). This age gradient may partly reflect the contribution of death, which was the largest component of the composite unsuccessful outcome (55.6%). Age-related comorbidities, frailty, treatment intolerance and delayed clinical recovery may also contribute to poorer outcomes among older patients. The lower odds observed among females (aOR = 0.65) may reflect differences in smoking exposure, occupational circumstances, healthcare-seeking behaviour, treatment adherence or comorbidity patterns, although these mechanisms were not directly assessed in TBIS.

Unemployment and no or primary education were associated with unsuccessful treatment, whereas tertiary education and healthcare-worker status were associated with lower adjusted odds. These findings highlight the potential importance of socioeconomic resources, health literacy, familiarity with healthcare systems, continuity of care and access to treatment support. The association with urban residence may reflect population mobility, competing socioeconomic demands, service workload or interruptions in continuity of care rather than an adverse effect of urban residence itself. Similarly, associations involving ethnicity and citizenship should be understood within their broader social, geographic and healthcare contexts and not interpreted as biological determinants. The higher adjusted odds among non-Malaysian patients (aOR = 2.15) were largely explained by transfers out: when transferred-out patients were excluded, the aOR fell to 0.88, and in the death-only model it was 1.22. The outcome definition therefore materially affected this association. The excess odds among non-Malaysian patients were driven mainly by transfers without a documented outcome: their odds of death were only modestly higher (aOR = 1.22; 95% CI: 1.14–1.31), and when transferred-out patients were excluded their odds of an unsuccessful outcome were lower than those of Malay patients (aOR = 0.88; 95% CI: 0.83–0.93). This finding should be read mainly as evidence of gaps in cross-facility continuity of care and outcome recording for mobile and migrant patients, rather than as a uniformly higher risk of poor outcomes in this group. By contrast, in a Korean national cohort, foreign residents had lower treatment success than Korean nationals, and this association persisted

when deaths were excluded [21]; differences in migrant populations, referral pathways and outcome definitions limit direct comparison. This reinforces the importance of accessible services, effective referral systems and continuity of treatment across facilities for mobile and migrant populations.

Among the clinical and behavioural factors, TB meningitis showed the strongest association with unsuccessful treatment (aOR = 5.15; 9.49 for death alone), consistent with its severity. HIV-positive status (aOR = 3.36) and an unrecorded HIV status, including not tested (aOR = 2.80), were also strongly associated, supporting continued integration of tuberculosis and HIV screening, treatment, and follow-up services. Having no DOT supervisor (aOR = 4.72) was strongly associated with unsuccessful treatment, although this field is recorded at the start of the intensive phase and may partly reflect early disengagement; it should therefore not be interpreted as a pre-treatment risk factor. Smoking was also associated with higher odds of an unsuccessful outcome, reinforcing the relevance of incorporating smoking-cessation support into tuberculosis care. Diabetes mellitus had a smaller but statistically significant association, consistent with a single-centre Philippine cohort in which comorbidity (diabetes, hypertension or HIV) was associated with unsuccessful treatment in unadjusted analysis [22], supporting appropriate monitoring and coordinated management of tuberculosis and diabetes. Absence of a recorded BCG scar was also associated with unsuccessful treatment; however, BCG-scar status represents a recorded physical finding and should not be interpreted as definitive evidence of vaccination history or as proof that BCG vaccination directly improves treatment outcomes.

The composite unsuccessful outcome included death, treatment failure, loss to follow-up and transfer without a documented successful outcome. These events may arise through different clinical, behavioural, and healthcare-system pathways. Studies also differ in how transferred-out patients are handled: two recent African cohorts excluded them from the outcome analysis [23,24], whereas we retained them in the unsuccessful composite and examined their exclusion in a sensitivity analysis, and Lee et al. separated mortality from failure and loss to follow-up by repeating their analysis without deaths [21]. Such differences limit direct comparison of treatment-success rates between studies. The identified factors should therefore be interpreted as markers of patients who may require additional clinical assessment or programme support rather than as evidence of a single causal pathway. Nevertheless, their consistency across the regression and SHAP analyses strengthens their relevance for population-level tuberculosis programme planning in Malaysia.

4.3. Machine-Learning Performance and Explainability

The nine models produced similar discrimination in the temporal test set, with ROC-AUC values ranging from 0.729 to 0.754. Gradient boosting, XGBoost, LightGBM and the stacked ensemble differed by 0.002 or less, so no model can be described as superior. Because unsuccessful treatment was designated as the positive class, the reported sensitivity, PPV and F1 score represent the classification of unsuccessful treatment outcomes; at the 0.20 threshold, about two-thirds of unsuccessful outcomes were identified at the cost of many false positives. Reporting performance for the event class is essential, because overall metrics can mask poor detection of unsuccessful outcomes: in a South African cohort of patients with drug-resistant tuberculosis and HIV, a decision-tree classifier with an accuracy of approximately 69% identified only 31% of unsuccessful outcomes (F1 score 0.43), whereas recall for successful outcomes was 92% [25].

The ROC-AUC values indicated moderate discrimination across all models. This finding reflects the complexity of tuberculosis treatment outcomes and the constraints of prediction based solely on routinely collected registry variables. Class imbalance, with unsuccessful outcomes accounting for 20.2% of the analytical cohort, also presents an inherent

challenge for prediction within national surveillance data. The findings therefore demonstrate that the available TBIS predictors contain meaningful predictive information while also indicating that treatment outcomes are influenced by additional clinical, behavioural, social, and healthcare-system factors not fully represented in the registry.

The conventional multivariable logistic-regression model, evaluated on the same test episodes, produced a ROC-AUC of 0.741, 0.012 lower than gradient boosting. This small difference supports the complementary rather than competing roles of the two analytical approaches. Logistic regression provided directly interpretable adjusted measures of association, whereas the machine-learning models evaluated multivariable predictive patterns and potential non-linear relationships. This observation is consistent with broader evidence showing that machine-learning methods do not invariably achieve substantially greater discrimination than logistic regression when both approaches use the same routinely collected predictors [14]. Comparable tuberculosis studies have also reported moderate discrimination, with performance varying according to the study population, outcome definition, predictor availability and validation approach [8–13].

SHAP analysis added an important interpretability component to the machine-learning framework by demonstrating the magnitude and direction of each predictor's contribution to the gradient-boosting output [19]. The broad agreement between the SHAP patterns and adjusted regression findings, particularly for age, sex, employment status, educational attainment, HIV status, chest radiography, BCG-scar status and DOT supervisor, strengthens confidence that the model captured clinically and programmatically relevant patterns. SHAP magnitude and adjusted odds ratios also answer different questions: TB meningitis had the largest adjusted odds ratio (aOR = 5.15) but a small mean absolute SHAP value (0.041) because it was recorded in only 1.4% of episodes and therefore changed few individual predictions, whereas common characteristics with moderate associations, such as ethnicity/citizenship and age, contributed more to predictions across the cohort. However, SHAP values explain the behaviour of the fitted model and do not establish causal relationships or represent the expected effect of modifying an individual predictor.

Taken together, the findings support a hybrid analytical framework that combines conventional epidemiological analysis with interpretable machine learning. The present model is most valuable for identifying population-level patterns and informing further development of registry-based risk-stratification strategies. External validation, prospective evaluation and a formal fairness evaluation would be required before considering its use for individual clinical decision-making or automated programme allocation.

4.4. Implications for Tuberculosis Programme and Healthcare Delivery

The finding that 20.2% of treatment episodes had an unsuccessful outcome indicates a continuing clinical and programme burden for tuberculosis services in Malaysia. The identified predictors can support a structured approach to recognising patients who may require enhanced assessment and follow-up. At treatment initiation and throughout treatment, older age, TB meningitis, HIV infection, unemployment, lower educational attainment, smoking, diabetes mellitus and absence of a BCG scar could prompt closer evaluation of clinical, behavioural, and social support needs. These characteristics should inform supportive care planning rather than be treated as deterministic classifications or used to restrict access to services.

The results reinforce the importance of coordinated, patient-centred tuberculosis care. Patients living with HIV may benefit from closer integration of tuberculosis and HIV services, including coordinated appointments, medication review and continuity of follow-up. Older patients may require additional assessment of comorbidities, treatment tolerance, and changes in clinical status. Smoking-cessation support and appropriate diabetes management

could be incorporated into routine tuberculosis care. Patients experiencing unemployment or other socioeconomic vulnerabilities may benefit from treatment navigation, flexible follow-up arrangements and referral to available social-support services. These approaches are consistent with the priorities of Malaysia's National Strategic Plan to End TB 2021–2030 [4].

The associations involving urban residence and non-Malaysian citizenship highlight the importance of continuity and equity within tuberculosis services. Urban programmes may need to address challenges related to population mobility, clinic workload, transport, fragmented care and movement between healthcare facilities. For mobile and migrant populations, among whom the excess risk was largely related to transfers out, clear referral procedures, effective transfer of treatment information, accessible follow-up and continuity across administrative or geographical boundaries may reduce interruptions in care. Ethnicity and citizenship should not be used independently to determine eligibility for services or intensity of treatment monitoring; instead, they should prompt assessment of the underlying social and healthcare barriers experienced by individual patients. Because sensitivity and calibration differed between subgroups, any programme use of the model would require a formal fairness evaluation and a decision on whether these inputs are acceptable.

The study also highlights the importance of strengthening TBIS data quality for surveillance, programme evaluation and future predictive applications. Only 138,231 of 193,862 episodes (71.3%) had complete predictor information, and ethnicity was blank for almost all non-citizen records. Standardised variable definitions, automated field validation, routine completeness audits, staff training and regular feedback to reporting facilities could improve the reliability and analytical value of the registry, consistent with WHO guidance on tuberculosis data generation and use [5]. Future registry development could also incorporate more detailed information on treatment adherence, microbiological response, drug resistance, comorbidity severity, healthcare access and relevant social circumstances.

The hybrid analytical framework used in this study offers a practical basis for translating national registry data into programme intelligence. Multivariable regression identifies independently associated factors, machine learning integrates multiple predictors into a unified risk pattern, and SHAP provides an interpretable explanation of model behaviour. Following appropriate temporal and external validation, this framework could support risk-informed monitoring, service planning, and resource prioritisation while complementing—not replacing—clinical judgement and equitable patient-centred care.

4.5. Strengths and Limitations

A major strength of this study was the use of a large national tuberculosis registry covering all Malaysian states and federal territories. The analytical cohort comprised 193,862 treatment episodes registered between 2014 and 2021, enabling demographic, socioeconomic, behavioural, and clinical factors to be examined under routine programme conditions. The integration of multivariable logistic regression, eight individual models, a stacked ensemble, a patient-grouped temporal test set, bootstrap confidence intervals, calibration and decision-curve analysis, subgroup and sensitivity analyses and SHAP analysis enabled adjusted associations, predictive performance and model behaviour to be examined within a unified, reproducible analytical framework. The study also addressed a programme-relevant outcome and was reported with reference to STROBE and TRIPOD+AI guidance [15,16].

Several limitations should be acknowledged. First, the retrospective registry design depended on the completeness and accuracy of routinely recorded information and was susceptible to coding errors, outcome misclassification, residual confounding and unmeasured variation in healthcare delivery. Missing values were common for some predictors (for example, employment status in 27,835 episodes and number of dependants in 11,157) and were retained as a separate 'Not recorded' category; associations were similar with multiple

imputation and complete-case analysis, but missingness related to unmeasured factors cannot be excluded. Registrations from 2022–2023 were excluded because follow-up was incomplete, and the registry patient identifier is close to one per episode, so re-registrations under a new identifier could not be linked.

Second, the composite unsuccessful outcome combined death, treatment failure, loss to follow-up, and transfer without a documented successful outcome. These events may arise through different clinical, behavioural and healthcare-system pathways, and deaths accounted for 55.6% of unsuccessful outcomes, so the composite model is driven mainly by mortality. Some patients classified as transferred out may also have completed treatment elsewhere without their final outcomes being recorded in the analysed registry extract. Discrimination was similar when transferred-out patients were excluded (ROC-AUC 0.744 for gradient boosting), but the association with non-Malaysian citizenship depended on the classification of transfers. Death and the other unsuccessful outcomes were also modelled separately, with full performance measures in Supplementary Table S4; separate models may suit different interventions, such as clinical review for patients at risk of death and adherence and continuity support for those at risk of disengagement.

Third, the prediction models were restricted to variables routinely available in TBIS. Potentially informative data on treatment adherence, microbiological response, drug resistance, disease severity, medication-related adverse events, comorbidity severity, mental health, social support, income and healthcare access were unavailable or not included. The absence of these variables may partly explain the moderate discrimination observed across the models. In addition, two predictors (HIV status and DOT supervisor) are recorded at or shortly after treatment initiation, and exact recording timestamps were not available, so temporal leakage cannot be excluded. Without them, ROC-AUC fell to 0.717 for logistic regression and 0.730 for gradient boosting; the difference from the main models (0.023 for gradient boosting) is an upper bound on the contribution of these fields, part of which may reflect information recorded after treatment initiation. Prospective use would require time-stamped recording of predictors.

Fourth, unsuccessful treatment outcomes represented 20.2% of the analytical cohort, resulting in an imbalanced outcome distribution. Although calibration, decision curves, subgroup performance and a temporal test set were examined, the models have not been externally validated, and performance differed between subgroups, with sensitivity ranging from 0.353 to 0.930 at the same threshold. Error rates also differed between subgroups, with gaps of 0.488 in sensitivity and 0.578 in the false-positive rate across ethnicity/citizenship groups (Supplementary Table S8), mainly because predicted risk differed between them. These subgroup estimates are descriptive; whether such differences are acceptable, or whether group-specific thresholds are needed, should be decided with programme stakeholders before any use. The 0.20 threshold is exploratory: it depends on an assumed cost ratio of four to one that has not been confirmed with programme stakeholders, and an operational threshold should be chosen according to the consequences of missed cases and the capacity available to support flagged patients; at 0.20, about 404 of every 1000 patients would be flagged. In addition, the temporal test period (2020–2021) coincided with the COVID-19 pandemic, which disrupted tuberculosis services and may have altered case mix, treatment supervision and outcome recording. Case mix and recording did differ between the development and test cohorts: occupation was not recorded for 18.2% of development but 1.6% of test episodes, and deaths made up 62.2% vs. 53.5% of unsuccessful outcomes (Supplementary Table S2). Calibration of the composite models was nevertheless preserved in the test set (O/E ratio 1.02 for gradient boosting), whereas the component models were not calibrated in the large, although updating their intercept on the most recent year largely corrected

this (Supplementary Table S9). Performance may differ in later periods. The findings therefore describe the performance of the models within the Malaysian registry and should not be interpreted as evidence of readiness for individual clinical deployment.

Individual-level TBIS data cannot be shared publicly because of national data-governance requirements. Nevertheless, the large national cohort, prespecified predictors, temporal test-set assessment and convergence between the multivariable and SHAP findings provide a strong foundation for population-level interpretation and independent validation of registry-based prediction strategies.

5. Conclusions

This national registry study of 193,862 tuberculosis treatment episodes found that 20.2% had an unsuccessful treatment outcome. TB meningitis, having no DOT supervisor, HIV infection or an unrecorded HIV status, retreatment, older age, male sex, unemployment, urban residence, smoking, diabetes mellitus and absence of a recorded BCG scar were associated with higher adjusted odds of an unsuccessful outcome. Non-Malaysian citizenship was also associated with higher adjusted odds, but this association was largely explained by transfers out and reversed when transferred-out patients were excluded (aOR = 0.88; 95% CI: 0.83–0.93). In contrast, healthcare-worker status, tertiary education and having a family member as DOT supervisor were associated with lower odds. These findings identify population groups who may benefit from enhanced clinical monitoring, integrated management and additional social or programme support.

Among the evaluated models, gradient boosting, XGBoost, LightGBM and the stacked ensemble achieved near-identical discrimination in the 2020–2021 temporal test set (ROC-AUC 0.752–0.754), compared with 0.741 for conventional logistic regression, with good calibration for the main models. Classification at the 0.20 threshold should be regarded as exploratory until an operational threshold is agreed with programme stakeholders. SHAP analysis improved the interpretability of the model and demonstrated patterns that were broadly consistent with the multivariable regression findings. The complementary use of conventional epidemiological analysis, machine learning, and SHAP therefore provides a practical framework for extracting clinically and programmatically relevant information from national tuberculosis registry data.

The moderate discrimination observed across the models highlights the complexity of predicting tuberculosis treatment outcomes using routinely collected and imbalanced registry data. The principal value of the present findings lies in population-level risk identification and tuberculosis programme planning. Strengthening TBIS data completeness and expanding the availability of relevant clinical, microbiological, adherence and social information could further enhance future predictive applications. External validation and fairness evaluation remain necessary before applying a registry-based prediction model to individual clinical decisions or automated resource allocation in Malaysian tuberculosis care.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/healthcare14223916/s1, Table S1: Study outcome and analytic coding of 17 predictors derived from 18 TBIS source fields; Table S2: Comparison of the development and temporal test cohorts; Table S3: Performance at alternative decision thresholds; Table S4: Performance of the component-specific outcome models; Table S5: Subgroup performance with 95% confidence intervals; Table S6: Multivariable logistic regression excluding transferred-out patients; Table S7: Performance of the 15-predictor baseline model; Table S8: Subgroup error rates and between-category gaps; Table S9: Temporal recalibration of the composite and component models.

Author Contributions: Conceptualization, S.A.S. and R.A.K.; methodology, D.B.A.; formal analysis, D.B.A. and M.S.M.P.; resources, M.I.B.M.; writing—original draft preparation, D.B.A.; writing—review and editing, D.B.A., S.A.S., M.I.B.M., R.A.K. and M.S.M.P.; supervision, S.A.S. and R.A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This retrospective secondary registry study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of Universiti Kebangsaan Malaysia (project code JEP-2022-493; reference UKM PPI/111/8/JEP-2022-493; 28 July 2022) and the Medical Research and Ethics Committee, Ministry of Health Malaysia (NMRR ID-22-02644-XU1 [IIR]; 22 February 2023).

Informed Consent Statement: The requirement for individual informed consent was waived by the approving ethics committees because this study involved the secondary analysis of de-identified registry data with no direct participant contact. No identifiable participant information is reported.

Data Availability Statement: The data supporting the findings of this study were obtained from the Malaysian National Tuberculosis Information System, Ministry of Health Malaysia. Restrictions apply to the availability of these data because they contain confidential patient-level health information and were used under authorisation for the present study. The data are therefore not publicly available. Access may be considered by the data custodian following a formal request and approval by the Ministry of Health Malaysia and the relevant ethics and governance authorities.

Acknowledgments: The authors acknowledge the Disease Control Division, Ministry of Health Malaysia, as the custodian of the Malaysian National Tuberculosis Information System, from which the data used in this study were obtained. The views expressed in this article are those of the authors and do not necessarily represent the official views or policies of the Ministry of Health Malaysia.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
aOR	Adjusted odds ratio
AUC	Area under the receiver operating characteristic curve
BCG	Bacille Calmette–Guérin
CI	Confidence interval
DOT	Directly observed treatment
GenAI	Generative artificial intelligence
HIV	Human immunodeficiency virus
LightGBM	Light Gradient Boosting Machine
ML	Machine learning
MOH	Ministry of Health Malaysia
MREC	Medical Research and Ethics Committee
NMRR	National Medical Research Register
NPV	Negative predictive value
PPV	Positive predictive value
PR-AUC	Area under the precision–recall curve
ROC	Receiver operating characteristic
SHAP	SHapley Additive exPlanations
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TB	Tuberculosis
TBIS	National Tuberculosis Information System
TRIPOD+AI	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis plus Artificial Intelligence
UKM	Universiti Kebangsaan Malaysia
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

References

1. World Health Organization. *Global Tuberculosis Report 2025*; World Health Organization: Geneva, Switzerland, 2025. Available online: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2025 (accessed on 13 September 2026).
2. World Health Organization. *The End TB Strategy*; World Health Organization: Geneva, Switzerland, 2015. Available online: https://www.who.int/publications/i/item/WHO-HTM-TB-2015.19 (accessed on 13 September 2026).
3. Ministry of Health Malaysia. *Annual Report 2023*; Ministry of Health Malaysia: Putrajaya, Malaysia, 2025. Available online: https://moh.gov.my/moh/resources/Penerbitan/Penerbitan%20Utama/MOH_Annual_Report_2023_updt_2Jan2025.pdf (accessed on 8 November 2026).
4. Ministry of Health Malaysia. *National Strategic Plan to End TB 2021-2030*; Ministry of Health Malaysia: Putrajaya, Malaysia, 2021. Available online: https://www.moh.gov.my/en/publications-and-reports/policies-act-policies-guide-lines/strategic-plan/list-of-strategic-plans/national-strategic-plan-to-end-tb-2021-2030 (accessed on 13 September 2026).
5. World Health Organization. *Consolidated Guidance on Tuberculosis Data Generation and Use: Module 1-Tuberculosis Surveillance*; World Health Organization: Geneva, Switzerland, 2024. Available online: https://www.who.int/publications/i/item/9789240075290 (accessed on 13 September 2026).
6. Liew, S.M.; Khoo, E.M.; Ho, B.K.; Lee, Y.K.; Mimi, O.; Fazlina, M.Y.; Asmah, R.; Lee, W.K.; Harmy, M.Y.; Chinna, K.; et al. Tuberculosis in Malaysia: Predictors of treatment outcomes in a national registry. *Int. J. Tuberc. Lung Dis.* **2015**, *19*, 764–771. [CrossRef] [PubMed]
7. Tok, P.S.K.; Liew, S.M.; Wong, L.P.; Razali, A.; Loganathan, T.; Chinna, K.; Ismail, N.; Abdul Kadir, N. Determinants of unsuccessful treatment outcomes and mortality among tuberculosis patients in Malaysia: A registry-based cohort study. *PLoS ONE* **2020**, *15*, e0231986. [CrossRef] [PubMed]
8. Zakaria, Y.S.; Ariffin, N.A.; Ahmad, A.; Rainis, R.; Muslim, A.M.; Wan Ibrahim, W.M.M. Optimizing tuberculosis treatment predictions: A comparative study of XGBoost with hyperparameter in Penang, Malaysia. *Sains Malays.* **2025**, *54*, 279–290. [CrossRef]
9. Rodrigues, M.M.S.; Barreto-Duarte, B.; Vinhaes, C.L.; Araújo-Pereira, M.; Fukutani, E.R.; Bergamaschi, K.B.; Kritski, A.; Cordeiro-Santos, M.; Rolla, V.C.; Sterling, T.R.; et al. Machine learning algorithms using national registry data to predict loss to follow-up during tuberculosis treatment. *BMC Public Health* **2024**, *24*, 1385. [CrossRef] [PubMed]
10. Kheirandish, M.; Catanzaro, D.; Crudu, V.; Zhang, S. Integrating landmark modeling framework and machine learning algorithms for dynamic prediction of tuberculosis treatment outcomes. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 900–908. [CrossRef] [PubMed]
11. Peetluk, L.S.; Rebeiro, P.F.; Ridolfi, F.M.; Andrade, B.B.; Cordeiro-Santos, M.; Kritski, A.; Durovni, B.; Calvacante, S.; Figueiredo, M.C.; Haas, D.W.; et al. A clinical prediction model for unsuccessful pulmonary tuberculosis treatment outcomes. *Clin. Infect. Dis.* **2022**, *74*, 973–982. [CrossRef] [PubMed]
12. Peetluk, L.S.; Ridolfi, F.M.; Rebeiro, P.F.; Liu, D.; Rolla, V.C.; Sterling, T.R. Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults. *BMJ Open* **2021**, *11*, e044687. [CrossRef] [PubMed]
13. Kamulegeya, R.; Nabatanzi, R.; Semugenze, D.; Mugala, F.; Takuwa, M.; Nasinghe, E.; Musinguzi, D.; Namiiro, S.; Katumba, A.; Ssengooba, W.; et al. Artificial intelligence and machine learning-based prediction of tuberculosis treatment failure: A systematic review and meta-analysis. *PLoS ONE* **2026**, *21*, e0355124. [CrossRef] [PubMed]
14. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **2019**, *110*, 12–22. [CrossRef] [PubMed]
15. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]
16. von Elm, E.; Altman, D.G.; Egger, M.; Pocock, S.J.; Gøtzsche, P.C.; Vandenbroucke, J.P.; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *Lancet* **2007**, *370*, 1453–1457. [CrossRef] [PubMed]
17. Wynants, L.; van Smeden, M.; McLernon, D.J.; Timmerman, D.; Steyerberg, E.W.; Van Calster, B. Three Myths about Risk Thresholds for Prediction Models. *BMC Med.* **2019**, *17*, 192. [CrossRef] [PubMed]
18. van den Goorbergh, R.; van Smeden, M.; Timmerman, D.; Van Calster, B. The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1525–1534. [CrossRef] [PubMed]
19. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
20. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]

21. Lee, J.Y.; Myong, J.-P.; Kim, Y.; Jeong, I.; Kim, J.; Sin, S.; Kwon, Y.; Song, C.; Joh, J.-S. Identifying Predictors of Unfavorable Treatment Outcomes in Tuberculosis Patients. *Int. J. Environ. Res. Public Health* **2024**, *21*, 1454. [CrossRef] [PubMed]
22. Hinay, A.A., Jr.; Mamalintaw, M.A.; Damasin, J.M.L.; Dilangalen, B.J.S.A.; Montinola, B.A.S.; Napinas, C.J.S.; Valiente, L.E.R.L.; Insular, N.L.C.; Parilla, A.J.D.; Cadotdot, N.M.T.; et al. Sociodemographic and Clinical Predictors of Tuberculosis and Unsuccessful Treatment Outcomes in Davao City, Philippines: A Retrospective Cohort Study. *Int. J. Environ. Res. Public Health* **2025**, *22*, 1154. [CrossRef] [PubMed]
23. Nxumalo, E.L.; Sineke, N.; Dlatu, N.; Apalata, T.; Faye, L.M. Treatment Outcomes of Tuberculosis in the Eastern Cape: Clinical and Socio-Demographic Predictors from Two Rural Clinics. *Int. J. Environ. Res. Public Health* **2025**, *22*, 1804. [CrossRef] [PubMed]
24. Tsegaye, B.; Bedewi, Z.; Asnake, S. The Treatment Outcomes of Tuberculosis Patients at Adare General Hospital, Southern Ethiopia (A Five-Year Retrospective Study). *Trop. Med. Infect. Dis.* **2024**, *9*, 262. [CrossRef] [PubMed]
25. Hosu, M.C.; Faye, L.M.; Apalata, T. Predicting Treatment Outcomes in Patients with Drug-Resistant Tuberculosis and Human Immunodeficiency Virus Coinfection, Using Supervised Machine Learning Algorithm. *Pathogens* **2024**, *13*, 923. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.