Infection and Drug Resistance

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Prediction of Tuberculosis Risk in the Elderly Population of Eastern China: Development and Validation of Multiple Machine Learning Models

Xiaofei Yu[1,*], Zhongqi Li[2,*], Hongmei Guo[3], Xiu Chen[4], Hui Jiang[5]

[1]Department of Stomatology, Zhongda Hospital, Southeast University, Nanjing, Jiangsu, People's Republic of China; [2]Anhui Province Clinical Research Center for Critical Respiratory Medicine, the First Affiliated Hospital of Wannan Medical College (Yijishan Hospital of Wannan Medical College), Wuhu, People's Republic of China; [3]Department of Chronic Communicable Disease, Disease Control and Prevention of Yangzhong City, Zhenjiang, Jiangsu Province, People's Republic of China; [4]Department of General Surgery, The First Affiliated Hospital with Nanjing Medical University, Nanjing, Jiangsu Province, People's Republic of China; [5]Department of Chronic Communicable Disease, Disease Control and Prevention of Zhenjiang City, Zhenjiang, Jiangsu Province, People's Republic of China

*These authors contributed equally to this work

Correspondence: Hui Jiang, Department of Chronic Communicable Disease, Disease Control and Prevention of Zhenjiang City, 9 Huangshan S Road, Zhenjiang, Jiangsu Province, 212000, People's Republic of China, Tel +86-15358599683, Fax +86-511-84434786, Email Zjcdcjh@sina.com; Xiu Chen, Department of General Surgery, The First Affiliated Hospital with Nanjing Medical University, No. 300 Guangzhou Road, Nanjing, Jiangsu Province, 210029, People's Republic of China, Email chenxiu@njmu.edu.cn

Background: Tuberculosis (TB) remains a significant public health burden among older adults, yet predictive tools for this population are limited. This study aimed to develop and validate machine learning models to predict TB risk among older adults in Eastern China.

Methods: A prospective cohort of 33,935 participants aged ≥60 years was followed for over 8 years. TB diagnosis was confirmed through linkage with the national TB surveillance system. LassoCox regression was used to identify key predictors of TB risk. Four machine learning models—CoxBoost, Generalized Boosted Models (GBM), LassoCox, and Random Survival Forests (RSF)—were developed and compared. Model performance was evaluated using time-dependent area under the receiver operating characteristic curve (AUC), Brier score, and concordance index.

Results: During follow-up, 387 participants developed TB, yielding an incidence rate of 134.5 per 100,000 person-years. The LassoCox model identified 14 predictors, including sex, alcohol consumption, dietary quality, body mass index, and C-reactive protein levels. Among the four models, the LassoCox model demonstrated the best discriminatory ability with an AUC of 0.717 (95% CI: 0.692–0.742), followed by GBM (AUC: 0.712, 95% CI: 0.687–0.737), CoxBoost (AUC: 0.708, 95% CI: 0.683–0.733), and RSF (AUC: 0.637, 95% CI: 0.611–0.663). The LassoCox model also demonstrated satisfactory calibration, with a Brier score of 0.338. Decision curve analysis confirmed clinical utility at threshold probabilities below 20%. Kaplan-Meier survival analysis showed significant differences between risk groups (log-rank $P < 0.001$), though survival curves revealed limited separation between low- and high-risk groups.

Conclusion: The LassoCox model demonstrated acceptable predictive performance for TB risk in older Chinese adults. These findings suggest that machine learning-based risk prediction tools could facilitate targeted TB screening by identifying high-risk individuals in aging populations, thereby enabling more efficient allocation of screening resources and earlier intervention. However, further model refinement and external validation in diverse populations are warranted before clinical implementation.

Plain Language Summary: Tuberculosis (TB) remains a serious health concern for older adults, particularly in countries like China. However, there are few tools available to predict who is most at risk. In this study, we followed over 33,000 adults aged 60 years and older in Eastern China for more than 8 years. We used computer-based methods called machine learning to build models that predict TB risk. Among the models tested, the LassoCox model performed best, correctly identifying individuals at higher risk. Key factors linked to TB risk included sex, alcohol use, diet quality, body weight, and markers of inflammation in the blood. Our findings suggest that these prediction tools could help doctors and public health officials identify older adults who might benefit most from TB screening. Further research is needed to improve these models and test them in other populations.

https://doi.org/10.2147/IDR.S587778
Received: 9 December 2025
Accepted: 6 March 2026
Published: 14 March 2026

Keywords: tuberculosis, elderly tuberculosis, incidence, risk factor, machine learning

Introduction

Tuberculosis (TB) continues to be a leading contributor to morbidity and mortality among older adults globally, with a particularly high burden in low- and middle-income countries. Older individuals, particularly those aged 65 years and above, are at increased risk due to immunosenescence, the presence of comorbid conditions (such as diabetes and chronic respiratory diseases), and malnutrition. These factors collectively enhance susceptibility to both TB infection and the progression to active disease.[1] Moreover, the elderly often exhibit atypical or nonspecific clinical manifestations of TB, including mild fever, fatigue, weight loss, and a persistent, non-specific cough, which frequently leads to misdiagnosis or delayed identification. Such delays not only heighten the risk of disease transmission but also contribute to poorer health outcomes in this vulnerable population.[2,3]

In the WHO regions of the Eastern Mediterranean, South-East Asia, and Western Pacific, the incidence of TB is notably higher among the elderly, with notification rates steadily increasing with advancing age. The highest burden of TB is observed in individuals aged 65 years and older.[4] Projections from the WHO suggest that the elderly population in China, defined as those aged 60 years and older, will grow substantially, from 12.4% (approximately 168 million people) in 2010 to 28.0% (approximately 402 million people) by 2040.[5] Early detection and prompt treatment are critical for improving TB-related outcomes in older adults, as studies have shown that timely intervention can significantly reduce both morbidity and mortality [6]. Despite this, TB screening remains underutilized in older populations, particularly in regions with high TB burden, where healthcare systems may not prioritize screening for this age group.[6] This highlights the growing need for targeted screening programs in aging populations, particularly in countries like China, where TB remains a significant public health threat.[7] China has implemented active screening strategies among older adults, including symptom assessment and chest radiography (CXR), to identify TB cases. However, given the large population size, it is not feasible to screen all elderly individuals. Thus, combining age with one or more TB risk factors could enable more targeted and efficient identification of high-risk individuals.

Despite the recognized need for early TB detection in older adults, current screening approaches predominantly rely on symptom-based assessments and chest radiography, which may lack sensitivity in individuals with atypical presentations. Moreover, existing risk prediction tools have primarily been developed for general adult populations and often fail to incorporate geriatric-specific risk factors such as nutritional status, physical function, and age-related comorbidities. There remains a significant gap in validated, prospective risk prediction models specifically designed for elderly populations in high-burden settings.

Previous studies have applied machine learning techniques to TB diagnosis and treatment outcome prediction with promising results, demonstrating that plasma immune profiling combined with ML algorithms could effectively distinguish active TB from latent infection and predict treatment response.[8,9] Similarly, ML approaches have been used to enhance the interpretation of chest radiographs and to predict drug resistance patterns. However, most existing ML-based TB risk prediction models have focused on diagnostic accuracy in symptomatic patients or treatment outcomes in confirmed cases, rather than on prospective identification of at-risk individuals in community settings. Furthermore, few studies have specifically targeted elderly populations, who present unique challenges due to atypical symptom presentation and the presence of multiple comorbidities.[3] To our knowledge, this is the first study to develop and compare multiple survival-based machine learning models for predicting incident TB risk specifically among community-dwelling older adults in a high-burden setting.

Materials and Methods

Study Population

As published previously,[10] this study targeted individuals aged 65 years and older residing in Zhenjiang City, Jiangsu Province, who received essential public health services between January and December 2016. These services, offered annually and free of charge as part of a national health initiative for the elderly, comprised both a structured questionnaire and a comprehensive clinical evaluation. The questionnaire captured demographic information and health-related

behaviors, while the clinical assessment included symptom screening and diagnostic testing such as fasting blood glucose, lipid profile, electrocardiography (ECG), complete blood count (CBC), urinalysis, abdominal ultrasound, and other pertinent examinations.

For participants exhibiting symptoms suggestive of tuberculosis—such as a persistent cough lasting more than two weeks, hemoptysis, blood-tinged sputum, unexplained weight loss, anorexia, fatigue, low-grade fever (particularly in the afternoon or evening), night sweats, chest discomfort, dyspnea, or lymphadenopathy—chest radiography (CXR) was conducted as a follow-up evaluation.

Outcome Definition

As tuberculosis is a notifiable disease in China, all confirmed cases are required to be registered in the Tuberculosis Management Information System. To identify incident cases of active tuberculosis, participant records were cross-referenced using unique personal identifiers across two data platforms: the Tuberculosis Management Information System and the National Basic Public Health Services Information System. Final confirmation of TB diagnosis was obtained through physician review at the designated local tuberculosis treatment center. Individuals with a pre-existing diagnosis of active TB at baseline were excluded from the analytic cohort.

Statistical Analysis

Descriptive statistics were used to summarize baseline characteristics. Categorical variables were expressed as frequencies and percentages, with group comparisons performed using the Pearson Chi-square test or Fisher's exact test, as appropriate. Continuous variables were summarized using the interquartile range (IQR) to represent the central tendency and variability of the data. The time to the onset of active tuberculosis was analyzed across diabetes status groups using Kaplan–Meier survival analysis. Cumulative incidence was compared between groups using the Log rank test. TB incidence was calculated as the number of new TB cases per 100,000 person-years. Person-years of follow-up for each participant were computed from the baseline visit to the earliest occurrence of active TB diagnosis, death, loss to follow-up, or the study endpoint, April 1, 2025. The 95% confidence intervals (CIs) for the incidence rates were derived using the Poisson distribution. Differences in TB incidence between groups were assessed using two-sample Poisson rate tests.

The data were randomly split into a training set (80% of the total sample) and a validation set (20% of the total sample) to develop and evaluate the predictive models. To identify key predictors of TB incidence, variable selection was performed using Lasso-Cox regression, implemented via the glmnet package. A 10-fold cross-validation approach was employed to determine the optimal penalty parameter (λ.min), retaining variables with non-zero coefficients. Four prognostic models were developed and compared: (1) CoxBoost: A gradient boosting model for Cox regression, with 200 boosting iterations and a penalty parameter of 500. (2) Gradient Boosting Machine (GBM): A tree-based Cox proportional hazards model, incorporating 300 trees, an interaction depth of 5, and a learning rate of 0.01. (3) Lasso-Cox: A sparse Cox regression model with L1 regularization ($\alpha = 1$), tuned using 10-fold cross-validation. (4) Random Survival Forest (RSF): An ensemble model consisting of 500 survival trees, using log-rank splitting for node division. For the LassoCox model, continuous variables were automatically standardized internally by the glmnet package (standardize = TRUE, the default setting) to ensure that the L1 penalty was applied uniformly across predictors with different scales. The standardization was performed within the cross-validation procedure using training data only, and the coefficients were returned on the original scale for interpretation. For tree-based models (Random Survival Forest and Gradient Boosting Machine), no explicit standardization was applied, as these algorithms are inherently invariant to monotonic transformations of predictor variables.

Model performance was evaluated in the validation set using the following metrics: (1) Time-dependent Area Under the Receiver Operating Characteristic Curve (AUC) at the median event time, calculated using the timeROC package. (2) Brier score, adjusted for censoring using Kaplan-Meier estimates. (3) Concordance index (C-index), calculated using the survival package.

Survival differences between high-risk and low-risk groups were assessed using Kaplan-Meier survival curves. Groups were stratified based on the median predicted risk from the Lasso-Cox model. The Log rank test was used to compare survival curves, and hazard ratios (HRs) with 95% CIs were calculated using Cox regression. Missing data were

addressed through complete case analysis. Participants with incomplete baseline information (N = 5,187) were excluded from the final analytic cohort prior to model development. All statistical analyses were performed using R version 4.3.1. Relevant packages included survival (v3.5–7), glmnet (v4.1–8), gbm (v2.1.9), and randomForestSRC (v3.2.0). Reproducibility was ensured by fixing random seeds (set.seed(123)). A two-tailed p-value of less than 0.05 was considered statistically significant.

Results

Baseline Characteristics

After excluding 5,187 individuals due to incomplete baseline information, the final study cohort consisted of 33,935 elderly participants. These were randomly divided into a training set (N = 27,148) and a validation set (N = 6,787) in a 4:1 ratio. Baseline characteristics were well-balanced between the two subsets (Table 1). The mean age was 77.0 years (standard deviation [SD] = 6.74), with a median age of 76.0 years (range: 65.0–111.0). Females accounted for 54.1% (N = 18,352) of the population. The mean body mass index (BMI) was 1.55 kg/m^2 (SD = 0.743), with a median value of 1.00 kg/m^2 (range: 0–3.00). Diabetes was diagnosed in 11.4% (N = 3,869) of participants.

Table 1 Demographic Characteristics of the 33,935 Participants, Overall and by Test and Train Set

	Test Set	Train Set	Total
	(N=27148)	(N=6787)	(N=33935)
Age (years)			
Mean (SD)	77.0 (6.73)	77.1 (6.78)	77.0 (6.74)
Median [Min, Max]	76.0 [65.0, 111]	76.0 [65.0, 106]	76.0 [65.0, 111]
Sex			
Female	14689 (54.1%)	3663 (54.0%)	18,352 (54.1%)
Male	12459 (45.9%)	3124 (46.0%)	15,583 (45.9%)
BMI (kg/m^2)			
Mean (SD)	1.55 (0.743)	1.55 (0.747)	1.55 (0.743)
Median [Min, Max]	1.00 [0, 3.00]	1.00 [0, 3.00]	1.00 [0, 3.00]
Body mass index, kg/m2			
Underweight	1139 (4.2)	265 (3.9)	1404 (4.1)
Normal	12935 (47.6)	3237 (47.7)	16,172 (47.4)
Overweight	10052 (37.0)	2528 (37.2)	12,580 (37.1)
Obese	3022 (11.1)	757 (11.2)	3779 (11.1)
Diabetes			
No	24044 (88.6%)	6022 (88.7%)	30,066 (88.6%)
Yes	3104 (11.4%)	765 (11.3%)	3869 (11.4%)
Exercise			
No	21568 (79.4%)	5440 (80.2%)	27,008 (79.6%)
Moderate	1888 (7.0%)	473 (7.0%)	2361 (7.0%)
High intensity	3692 (13.6%)	874 (12.9%)	4566 (13.5%)
Diet			
Bad	25715 (94.7%)	6417 (94.5%)	32,132 (94.7%)
General	1221 (4.5%)	317 (4.7%)	1538 (4.5%)
Good	179 (0.7%)	43 (0.6%)	222 (0.7%)
Very good	33 (0.1%)	10 (0.1%)	43 (0.1%)
Smoking			
Never	22649 (83.4%)	5699 (84.0%)	28,348 (83.5%)
Quit smoking	324 (1.2%)	96 (1.4%)	420 (1.2%)
Currently smoking	4175 (15.4%)	992 (14.6%)	5167 (15.2%)

(*Continued*)

Table 1 (Continued).

	Test Set	Train Set	Total
	(N=27148)	(N=6787)	(N=33935)
Drinking			
Never	22767 (83.9%)	5696 (83.9%)	28,463 (83.9%)
Quit drinking	65 (0.2%)	16 (0.2%)	81 (0.2%)
Currently drinking	4316 (15.9%)	1075 (15.8%)	5391 (15.9%)
Time of following-up (years)			
Mean (SD)	8.47 (0.565)	8.48 (0.545)	8.47 (0.561)
Median [Min, Max]	8.49 [0.0110, 10.3]	8.49 [0.0110, 9.64]	8.49 [0.0110, 10.3]
Active tuberculosis			
No	26830 (98.8%)	6718 (99.0%)	33,548 (98.9%)
Yes	318 (1.2%)	69 (1.0%)	387 (1.1%)

A substantial proportion of individuals reported physical inactivity (79.6%, N = 27,008) and poor dietary quality (94.7%, N = 32,132). Regarding lifestyle factors, 83.5% (N = 28,348) had never smoked, and 83.9% (N = 28,463) had never consumed alcohol. The mean follow-up duration was 8.47 years (SD = 0.561), with a median of 8.49 years (range: 0.011–10.3). A total of 387 cases of active tuberculosis (TB) were identified, corresponding to an overall incidence rate of 1.1%, with 318 cases in the training set (1.2%) and 69 cases in the validation set (1.0%) (Table 1).

Incidence of Active Tuberculosis

Over a follow-up period exceeding eight years, 387 participants developed active TB, yielding an incidence rate of 134.5 per 100,000 person-years (95% confidence interval [CI]: 121.1–147.9). Within the training cohort, 318 TB cases were documented (138.1 per 100,000 person-years; 95% CI: 123.0–153.2), while 69 cases occurred in the validation cohort (120.0 per 100,000 person-years; 95% CI: 91.7–148.3). The median time to TB onset was 2,542 days (approximately 7 years), ranging from 485 to 3,365 days, reflecting variability in latency periods.

Correlation Analysis and Variable Importance in the Lasso-Cox Model

A correlation heatmap (Supplementary Figure 1) highlighted key associations among variables. The strongest positive correlation was observed between drinking and smoking status (r = 0.49), followed by smoking and sex (r = 0.46). Most other correlations were minimal (absolute value of r< 0.1), suggesting low multicollinearity among predictors and supporting the suitability of Lasso regression.

The Lasso-Cox regression model identified several key predictors for incident TB. Occasional alcohol consumption (Drinking1) had the strongest positive association, with an estimated coefficient of approximately 0.8. Female sex (Sex0) was associated with a lower risk, with a coefficient estimated at –0.7. Other influential variables included good dietary quality (Diet2, coefficient approximately –0.5), fair dietary quality (Diet1, approximately 0.3), BMI (approximately –0.4), impaired motor function (approximately 0.3), CRP (approximately 0.2), and moderate physical activity (Exercise1, approximately 0.2). Variables such as smoking status, WBC count, platelet count, hemoglobin, electrocardiographic abnormalities, age, and high levels of exercise were also retained in the model but had smaller coefficients, indicating comparatively limited impact on TB risk (Supplementary Figure 2).

Comparative Performance of Predictive Models

To assess model performance, ROC curves were generated for four prognostic algorithms: Lasso-Cox, GBM, CoxBoost, and RSF. The Lasso-Cox model exhibited the highest discriminatory performance with an AUC of 0.717 (95% CI: 0.639–0.790), followed closely by GBM (AUC = 0.712; 95% CI: 0.626–0.787) and CoxBoost (AUC = 0.708; 95% CI: 0.623–0.778). The RSF model yielded the lowest AUC at 0.637 (95% CI: 0.527–0.722) (Figure 1).

https://doi.org/10.2147/IDR.S587778

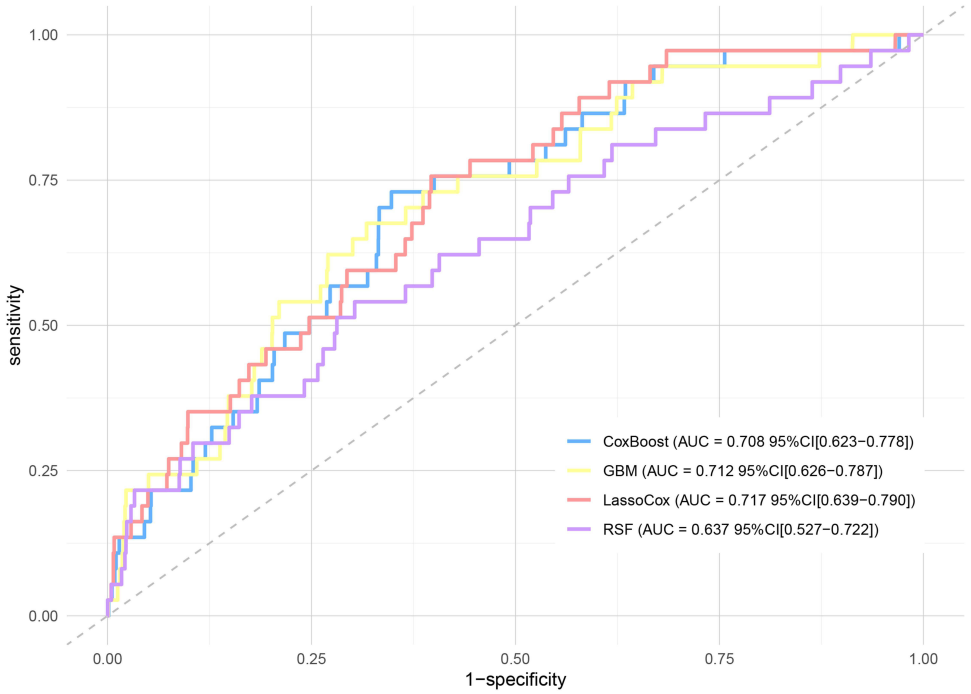

Figure 1 Receiving operating characteristic curve analysis for distinguishing between TB and non-TB CoxBoost, GBM, LassoCox, and RSF Models.

With regard to calibration, the Lasso-Cox model achieved the lowest Brier score (0.338) and the most favorable concordance index (0.299), suggesting superior calibration and risk stratification. GBM and CoxBoost models reported Brier scores of 0.356 and 0.326 and concordance indices of 0.307 and 0.306, respectively. The RSF model performed least effectively, with a Brier score of 0.346 and a concordance index of 0.361 (Figure 2). The Lasso-Cox model demonstrated the best discriminative performance in the validation set with a C-index of 0.702 (95% CI: 0.647–0.757), indicating good predictive ability. This model also showed the most stable performance between training and validation sets, suggesting minimal overfitting. The RSF model showed signs of overfitting with a training C-index of 0.989 but a validation C-index of only 0.624 (Supplementary Table 1).

Time-dependent AUC plots corroborated these findings. The Lasso-Cox curve consistently remained closer to the upper-left corner of the ROC space, reflecting superior sensitivity and specificity. GBM and CoxBoost displayed comparable performance, while the RSF model showed a notably flatter trajectory, consistent with its lower overall accuracy (Figure 3).

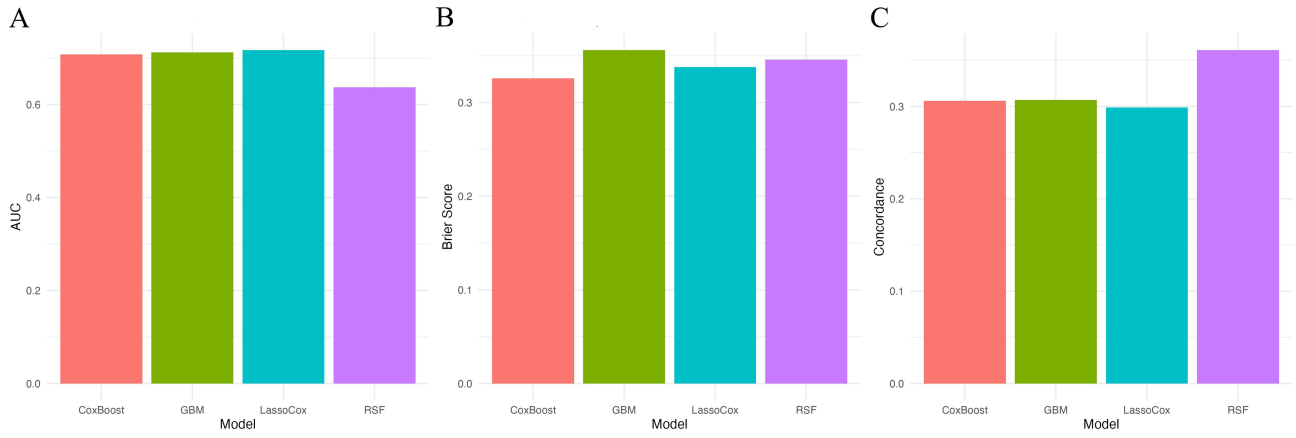

Figure 2 Performance Comparison of CoxBoost, GBM, LassoCox, and RSF Models Using AUC, Brier Score, and Concordance Index. (**A**) Time-dependent area under the receiver operating characteristic curve (AUC) at the median event time. (**B**) Brier score indicating calibration performance. (**C**) Concordance index (C-index) reflecting discriminatory ability.

https://doi.org/10.2147/IDR.S587778

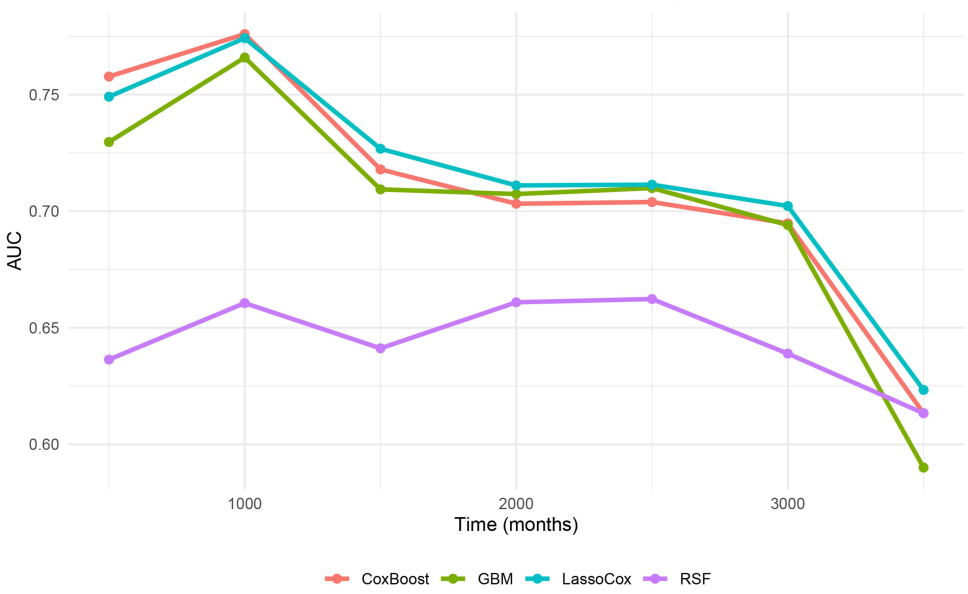

Figure 3 Time-Dependent AUC Curves for CoxBoost, GBM, LassoCox, and RSF Models.

Risk Stratification and Survival Analysis

Kaplan–Meier survival curves were used to evaluate the model's stratification capability. Participants were classified into low- and high-risk groups based on Lasso-Cox–derived risk scores. In the training set (N = 27,148), the low-risk group (N = 13,574) exhibited consistently higher survival probabilities compared to the high-risk group over the follow-up period. At 3,000 days, 13,269 and 13,110 participants remained at risk in the low- and high-risk groups, respectively. Similar trends were observed in the validation cohort, supporting the model's generalizability and utility in individualized risk prediction for active TB (Figure 4).

Risk Stratification and Sensitivity Analyses

To explore clinically meaningful risk thresholds beyond median-based classification, we performed additional stratification analyses using tertiles, quartiles, and fixed percentile cutoffs. All threshold values were derived from the training cohort and applied to the validation cohort to ensure unbiased external validation.

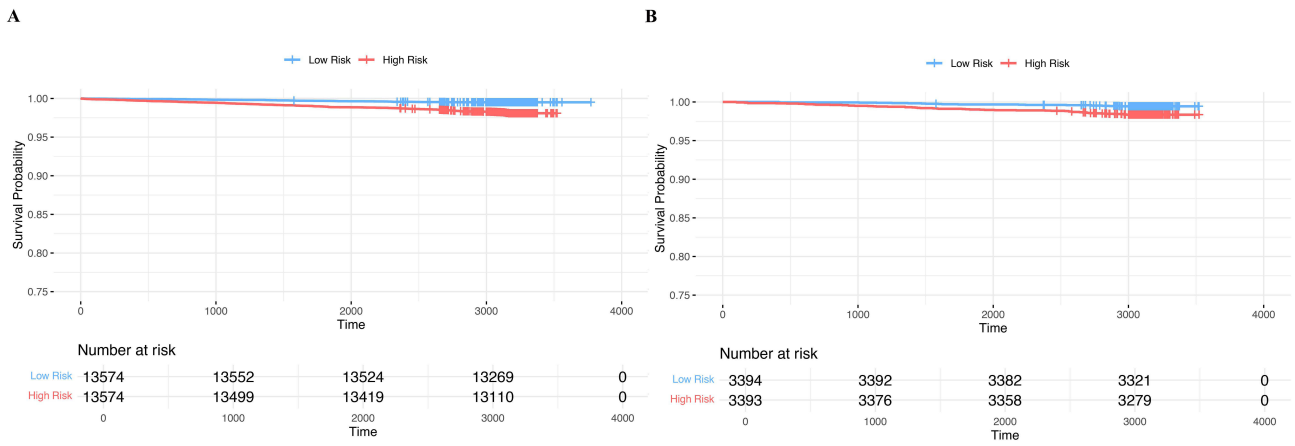

Figure 4 Kaplan–Meier overall survival curves with Different Risks Stratified by the LassoCox Model. (**A**) Survival curves in the training cohort. (**B**) Survival curves in the validation cohort. The solid blue line represents participants in the low-risk group, while the solid red line represents participants in the high-risk group.

https://doi.org/10.2147/IDR.S587778

In the tertile-based analysis, TB incidence rates were 50.6, 100.4, and 257.2 per 100,000 person-years for low-, intermediate-, and high-risk groups in the training cohort, respectively (log-rank $P < 0.001$). The high-risk tertile demonstrated a 5.09-fold increased TB risk compared to the low-risk group (HR = 5.09; 95% CI: 3.61–7.17). When applying the same cutoffs to the validation cohort, incidence rates were 36.2, 118.1, and 240.3 per 100,000 person-years, with a hazard ratio of 6.61 (95% CI: 2.98–14.66) for the high-risk versus low-risk tertile ($P < 0.001$).

Quartile-based stratification revealed a significant dose-response relationship. In the training cohort, TB incidence increased progressively from 36.3 per 100,000 person-years in Q1 to 78.1 in Q2, 146.2 in Q3, and 283.9 in Q4 (P for trend $< 2 \times 10^{-16}$). Compared to Q1, participants in Q4 had a 7.82-fold increased TB risk (HR = 7.82; 95% CI: 4.96–12.31). The validation cohort demonstrated consistent patterns, with incidence rates of 27.5, 102.9, 119.9, and 274.4 per 100,000 person-years across quartiles (P for trend < 0.001). The hazard ratio for Q4 versus Q1 was 9.97 (95% CI: 3.56–27.89).

Fixed threshold analyses demonstrated that identifying the top 10%, 20%, or 30% of participants as high-risk yielded consistently elevated hazard ratios in both training (range: 3.21–3.42) and validation (range: 3.17–3.79) cohorts (all $P <$ 0.001), supporting the model's utility for targeted screening strategies at various resource allocation levels (Supplementary Table 2).

To assess model robustness across different prediction horizons, we conducted sensitivity analyses at 1-year, 3-year, and 5-year time points (Supplementary Table 3). For short-term prediction (1-year), all models demonstrated excellent performance, with CoxBoost achieving the highest C-index of 0.917 (95% CI: 0.806–1.000), followed by Lasso-Cox (0.899) and GBM (0.853). At 3 years, CoxBoost and Lasso-Cox showed comparable performance (C-index: 0.763). For long-term prediction (5-year), GBM achieved the highest C-index of 0.710 (95% CI: 0.639–0.781), with all models maintaining values above 0.69.The gradual decrease in performance from 1-year to 5-year is consistent with inherent challenges of longer prediction horizons. Lasso-Cox demonstrated the most stable performance across all time points.

Discussion

In this longitudinal cohort analysis, we employed the Lasso-Cox regression technique to identify 14 key predictors associated with TB risk among older adults. Based on these variables, we constructed four ML models, among which the Lasso-Cox model demonstrated the most favorable performance in terms of both discrimination and clinical applicability, with the highest AUC.

A significant outcome of our study was the relatively elevated incidence of TB among elderly individuals in Zhenjiang City, Jiangsu Province, observed over an eight-year follow-up period. When comparing this rate with those from other countries, a stark contrast becomes evident, especially between regions with varying TB burdens. For example, literature reports indicate that the annual TB incidence among individuals aged 65 and over stands at approximately 10.9 per 100,000 in the United States and 11.2 per 100,000 in Germany.[11,12] Conversely, in countries with a higher TB burden, such as India, reported notification rates reach 305 per 100,000 for men and 81 per 100,000 for women.[13] Similarly, in South Africa, TB incidence rates among the elderly range from 518 to 684 per 100,000 for men and 193 to 314 per 100,000 for women.[14]

In contrast to these figures, our cohort showed a TB incidence rate of 134.5 per 100,000 person-years, which, while higher than some low-burden nations, remains markedly lower than those reported in other high-burden contexts. For instance, a comparable study in Taiwan with an eight-year follow-up documented an incidence rate of 175.5 per 100,000,[15] and a separate two-year investigation by Jun Cheng et al (2013) revealed a much higher rate of 481.8 per 100,000.[16] These disparities suggest that local healthcare systems, TB control policies, population characteristics, and screening practices may substantially influence disease burden, underscoring the need for context-specific strategies in TB prevention for older adults.

A particular challenge in managing TB in older populations is the often atypical and insidious onset of symptoms. Unlike younger patients, elderly individuals frequently present with non-specific manifestations—such as fatigue, low-grade fever, and unintentional weight loss—that are easily confused with other age-associated conditions or overlooked altogether.[6,8] This can result in delayed diagnosis and treatment, a pattern observed in our cohort where the median time

to TB onset was approximately seven years. Such delays can have serious consequences, increasing the risk of complications and transmission, especially in institutional settings like long-term care facilities.

Addressing these challenges requires early and accurate identification of at-risk individuals. Our findings support the integration of predictive modeling into public health interventions. Among the ML models we tested, the Lasso-Cox regression showed good predictive accuracy (AUC = 0.717), aligning with previous research that has validated machine learning approaches for enhancing TB diagnosis and prognosis.[9,17] Such models offer a data-driven method to stratify risk and allocate screening resources more efficiently.

In addition to clinical predictors, our results emphasize the significant role of behavioral and metabolic risk factors in TB development among the elderly. Lifestyle-related factors such as physical inactivity, inadequate nutrition, and even occasional alcohol intake were identified as independent risk factors. Although casual alcohol use may appear benign, existing evidence suggests that it can impair immune defenses and heighten susceptibility to infectious diseases, including TB.[18,19]

We also observed that comorbidities such as diabetes and elevated CRP levels were associated with increased TB risk. This is consistent with the literature, which highlights the immunosuppressive effects of chronic metabolic conditions and systemic inflammation.[20,21] Lönnroth et al (2013), for instance, emphasized the need for integrated care models that address both communicable and non-communicable diseases to mitigate TB risk.[20]

These findings advocate for a comprehensive public health strategy that goes beyond traditional TB control measures. Efforts to improve physical fitness, dietary quality, metabolic control, and moderation of alcohol use should be woven into TB prevention programs targeting older adults. Previous studies have demonstrated the effectiveness of such holistic interventions in reducing TB incidence and enhancing overall geriatric health outcomes.[22,23]

This study has several strengths. First, we utilized a large prospective cohort with a long follow-up period of over 8 years, which enhanced the reliability of our findings. Second, we compared multiple machine learning algorithms and provided comprehensive model evaluation using various performance metrics. Third, the use of decision curve analysis allowed us to assess the clinical utility of the prediction models.

Nonetheless, this study has several limitations that should be acknowledged. First, the cohort was geographically restricted to Zhenjiang City, which may affect the generalizability of our findings to other regions with different socioeconomic or health infrastructure contexts. Second, our reliance on the Zhenjiang CDC registry for follow-up data means that TB cases occurring outside this jurisdiction—particularly among participants who migrated—may have gone undetected, potentially underestimating TB incidence. Third, the AUC of 0.717, while acceptable, indicates that there is room for improvement in predictive accuracy, and future studies should consider incorporating additional predictors such as chest radiography findings, interferon-gamma release assay results, or socioeconomic factors. Fourth, while ML-based tools show great promise, challenges remain in terms of interpretability, data standardization, and integration into real-world health systems.

Nonetheless, this study has several limitations that should be acknowledged. First, the cohort was geographically restricted to Zhenjiang City, which may limit generalizability to other regions with different socioeconomic or health infrastructure contexts. Second, our reliance on the Zhenjiang CDC registry means that TB cases occurring outside this jurisdiction may have gone undetected, potentially underestimating TB incidence. Third, the AUC of 0.717, while acceptable, indicates room for improvement, and future studies should incorporate additional predictors such as chest radiography findings or interferon-gamma release assay results. Fourth, complete case analysis resulted in the exclusion of 5187 participants with incomplete baseline information. Individuals with missing data may represent a subpopulation with distinct characteristics, such as those with more severe illness or lower health literacy. Future studies should consider multiple imputation techniques to minimize potential selection bias. Fifth, predictor variables including dietary quality, physical activity, smoking, and alcohol consumption were self-reported, which may be subject to recall and social desirability bias. Future studies should incorporate objective measures such as accelerometry or biomarkers to validate self-reported data. Sixth, although internal validation was performed using a training-validation split, external validation in independent cohorts was not conducted. Multicenter studies are needed to establish generalizability before clinical implementation.

Conclusion

In this prospective cohort study of elderly adults in Eastern China, we developed and validated machine learning models for predicting TB risk. The LassoCox model demonstrated the best predictive performance with an AUC of 0.717, with key predictors including sex, alcohol consumption, dietary quality, BMI, and inflammatory markers. This model could be integrated into China's national basic public health service framework during routine annual health assessments, enabling risk stratification to prioritize high-risk individuals for chest radiography and early TB detection.

Abbreviations

ALT, alanine aminotransferase; AST, aspartate aminotransferase; AUC, area under the receiver operating characteristic curve; BMI, body mass index; CBC, complete blood count; CI, confidence interval; CRP, C-reactive protein; CXR, chest radiography; ECG, electrocardiography; GBM, generalized boosted models; HR, hazard ratio; IQR, interquartile range; ML, machine learning; RSF, random survival forests; SD, standard deviation; TB, tuberculosis; WBC, white blood cell count; WHO, World Health Organization.

Data Sharing Statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author, Dr. Hui Jiang (Email: Zjcdcjh@sina.com), upon reasonable request.

Ethics Approval and Informed Consent

The research protocol was evaluated and approved by the Ethics Committee of the Zhenjiang City Center for Disease Control and Prevention (Approval No. [2023] 005). All participants provided written informed consent prior to their inclusion in the study. This study was conducted in accordance with the Declaration of Helsinki.

Acknowledgments

The authors thank all investigators from the Jiangsu Provincial Center for Disease Control and Prevention, Center for Disease Control and Prevention of Zhenjiang City.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The work was supported by the the Provincial Clinical Specialty Capacity Improvement Project of the Department of Stomatology, Zhongda Hospital Affiliated to Southeast University (CZXM-ZK-54), and the Talent Introduction of the First Affiliated Hospital of Wannan Medical College (KY2960YR2530).

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Olmo-Fontánez AM, Turner J. Tuberculosis in an aging world. *Pathogens*. 2022;11(10):1101. doi:10.3390/pathogens11101101
2. Raghu S. Challenges in treating tuberculosis in the elderly population in tertiary institute. *Indian J Tuberc*. 2022;69(Suppl 2):S225–s31. doi:10.1016/j.ijtb.2022.10.008
3. Abbara A, Collin SM, Kon OM, et al. Time to diagnosis of tuberculosis is greater in older patients: a retrospective cohort review. *ERJ Open Res*. 2019;5(4):00228–2018. doi:10.1183/23120541.00228-2018
4. Caraux-Paz P, Diamantis S, de Wazières B, Gallien S. Tuberculosis in the Elderly. *J Clin Med*. 2021;10(24):5888. doi:10.3390/jcm10245888
5. World Health Organization. *World Health Organization China Country Assessment Report on Ageing and Health*. 2015

6. Wingfield T. Ending tuberculosis in older people: new strategies for an age-old disease. *Clinical Infectious Diseases*. 2023;77(10):1476.
7. Li J, Chung P-H, Leung CL, Nishikiori N, Chan EY, Yeoh E-K. The strategic framework of tuberculosis control and prevention in the elderly: a scoping review towards End TB targets. *Infect Diseases Poverty*. 2017;6(03):16–27. doi:10.1186/s40249-017-0284-4
8. Theodosiou AA, Read RC. Artificial intelligence, machine learning and deep learning: potential resources for the infection clinician. *J Infect*. 2023;87(4):287–294. doi:10.1016/j.jinf.2023.07.006
9. Yao F, Zhang R, Lin Q, et al. Plasma immune profiling combined with machine learning contributes to diagnosis and prognosis of active pulmonary tuberculosis. *Emerg Microbes Infect*. 2024;13(1):2370399. doi:10.1080/22221751.2024.2370399
10. Jiang H, Chen X, Lv J, et al. Prospective cohort study on tuberculosis incidence and risk factors in the elderly population of eastern China. *Heliyon*. 2024;10(3):e24507. doi:10.1016/j.heliyon.2024.e24507
11. Hochberg NS, Horsburgh JCR. Prevention of tuberculosis in older adults in the United States: obstacles and opportunities. *Clinl Infect Dis*. 2013;56(9):1240–1247. doi:10.1093/cid/cit027
12. Hauer B, Brodhun B, Altmann D, Fiebig L, Loddenkemper R, Haas W. Tuberculosis in the elderly in Germany. *Eur Respir J*. 2011;38(2):467–470. doi:10.1183/09031936.00199910
13. Yew WW, Yoshiyama T, Leung CC, Chan DP. Epidemiological, clinical and mechanistic perspectives of tuberculosis in older people. *Respirology*. 2018;23(6):567–575. doi:10.1111/resp.13303
14. Nanoo A, Izu A, Ismail NA, et al. Nationwide and regional incidence of microbiologically confirmed pulmonary tuberculosis in South Africa, 2004–12: a time series analysis. *Lancet Infect Dis*. 2015;15(9):1066–1076. doi:10.1016/S1473-3099(15)00147-4
15. Yen Y-F, Pan S-W, VY-F S, Chuang P-H, Feng J-Y, Su W-J. Influenza vaccination and incident tuberculosis among elderly persons, Taiwan. *Emerging Infect Dis*. 2018;24(3):498. doi:10.3201/eid2403.152071
16. Cheng J, Sun Y-N, Zhang C-Y, et al. Incidence and risk factors of tuberculosis among the elderly population in China: a prospective cohort study. *Infect Diseases Poverty*. 2020;9(1):64–76. doi:10.1186/s40249-019-0614-9
17. Li J, Chung PH, Leung CLK, Nishikiori N, Chan EYY, Yeoh EK. The strategic framework of tuberculosis control and prevention in the elderly: a scoping review towards End TB targets. *Infect Dis Poverty*. 2017;6(1):70.
18. Kong W, Sheng W, Zheng Y. Modification of the association between coffee consumption and constipation by alcohol drinking: a cross-sectional analysis of NHANES 2007-2010. *PLoS One*. 2024;19(10):e0311916. doi:10.1371/journal.pone.0311916
19. Ragan EJ, Kleinman MB, Sweigart B, et al. The impact of alcohol use on tuberculosis treatment outcomes: a systematic review and meta-analysis. *Int J Tuberc Lung Dis*. 2020;24(1):73–82. doi:10.5588/ijtld.19.0080
20. Foo C, Shrestha P, Wang L, et al. Integrating tuberculosis and noncommunicable diseases care in low- and middle-income countries (LMICs): a systematic review. *PLoS Med*. 2022;19(1):e1003899. doi:10.1371/journal.pmed.1003899
21. Boadu AA, Yeboah-Manu M, Osei-Wusu S, Yeboah-Manu D. Tuberculosis and diabetes mellitus: the complexity of the comorbid interactions. *Inter J Infect Dis*. 2024;146:107140. doi:10.1016/j.ijid.2024.107140
22. Yoo JE, Kim D, Choi H, et al. Anemia, sarcopenia, physical activity, and the risk of tuberculosis in the older population: a nationwide cohort study. *Therap Advn Chronic Dis*. 2021;12:20406223211015959. doi:10.1177/20406223211015959
23. Wagnew F, Gray D, Tsheten T, Kelly M, Clements ACA, Alene KA. Effectiveness of nutritional support to improve treatment adherence in patients with tuberculosis: a systematic review. *Nutr Rev*. 2024;82(9):1216–1225. doi:10.1093/nutrit/nuad120