

Original Article

Predictive nomogram for risk of pulmonary infection in lung cancer patients undergoing radiochemotherapy: development and performance evaluation

Yujie Huang, Guang Han

Department of Radiation Oncology, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430079, Hubei, China

Received November 14, 2024; Accepted January 15, 2025; Epub February 15, 2025; Published February 28, 2025

Abstract: Objective: To develop an accurate predictive model for identifying patients at high risk of pulmonary infection during radiochemotherapy. Methods: We retrospectively analyzed data from 544 lung cancer patients treated at Hubei Cancer Hospital between May 2019 and October 2022. The patients were divided into training and validation groups (7:3 ratio). An external validation cohort of 100 patients treated from November 2022 to January 2024 was also included. Feature selection and model development were performed using machine learning algorithms, including Lasso regression, Random Forest, XGBoost, and Support Vector Machine (SVM). Model performance was evaluated using Receiver Operating Characteristic (ROC) curves, calibration curves, and decision curve analysis. Results: Key predictive factors for pulmonary infection risk were identified, including diabetes, chronic obstructive pulmonary disease, chemotherapy intensity, chemotherapy cycles, antibiotic use, age, Karnofsky Performance Status score, systemic inflammation index, prognostic nutritional index, and C-reactive protein. A nomogram-based prediction model was constructed, achieving ROC curve Area Under the Curve values of 0.889 in the training set, 0.897 in the validation set, and 0.875 in the external validation set, demonstrating strong classification ability and stability. Conclusion: We developed a robust nomogram-based model incorporating eight key factors to predict the risk of pulmonary infection in lung cancer patients undergoing radiochemotherapy. This model can assist clinicians in early identification of high-risk patients, enabling timely interventions to improve patient outcomes and quality of life.

Keywords: Lung cancer, radiochemotherapy, pulmonary infection, nomogram, predictive model

Introduction

Lung cancer remains the leading cause of cancer-related morbidity and mortality globally and in China [1]. In recent years, its incidence has steadily increased, largely due to industrialization, environmental pollution, and high tobacco use [2]. Lung cancer is the most common cause of cancer-related deaths among middle-aged and elderly men in China, with its incidence in women ranking second only to breast cancer. Given this trend, primary lung cancer will continue to be a major focus of prevention, detection, and treatment efforts in China [3].

Current treatment options for lung cancer include surgery, chemotherapy, immunotherapy, radiotherapy, and molecular targeted thera-

pies [4]. For patients with advanced non-small cell lung cancer (NSCLC), radiochemotherapy is a key treatment modality, particularly for those who are ineligible for surgery or at high surgical risk [5]. While radiochemotherapy can significantly shrink tumors, improve local control, and extend survival, it also offers advantages over surgery, such as pain reduction and improved quality of life for advanced-stage patients [6]. However, one major limitation of radiochemotherapy is the increased risk of pulmonary infections [7]. The treatment can impair immune function and disrupt the respiratory defense mechanisms, making patients more vulnerable to infections [8]. Previous studies have identified several factors influencing pulmonary infection risk in these patients, such as age, comorbidities (e.g., COPD, diabetes), che-

Nomogram for predicting lung infection risk

motherapy intensity and cycle length, nutritional status, and inflammatory markers [7-9]. Pulmonary infections not only prolong hospitalization and increase medical costs but also severely affect patients' prognosis and quality of life [9]. Thus, an accurate predictive model for pulmonary infection risk is crucial for early intervention and infection prevention.

In recent years, machine learning technologies have gained widespread use in the medical field, providing new tools for clinical decision-making and personalized treatment [10-14]. Machine learning can analyze large volumes of clinical data, uncovering complex patterns that traditional statistical methods may miss, and processing multidimensional data, such as gene expression [12], imaging data, and clinical records [13]. Algorithms like Lasso regression, XGBoost, and Support Vector Machine (SVM) have been extensively applied in cancer prognosis, disease diagnosis, and treatment response prediction [14], helping to automatically identify and select the most predictive features to construct efficient models. Nomograms, built using regression analysis, simplify complex predictive models into visual charts that allow clinicians to easily calculate individual risk scores and assess the likelihood of diseases or adverse events [15-17]. By integrating multiple predictive variables, nomograms provide personalized risk assessments, offering valuable clinical utility to guide treatment decisions.

This study aims to identify potential factors influencing the risk of pulmonary infection, construct a nomogram prediction model, and evaluate its performance. We will apply various machine learning techniques for feature selection and model development, and comprehensively assess the model's accuracy using ROC curves, calibration curves, and decision curve analysis. Through this study, we aim to provide clinicians with an efficient and practical tool to make more precise treatment decisions in complex clinical settings, thereby improving patient prognosis and quality of life.

Methods and materials

Ethical statement

This study was approved by the Medical Ethics Committee of Hubei Cancer Hospital.

Clinical data

We retrospectively analyzed data from 544 lung cancer patients treated at Hubei Cancer Hospital between May 2019 and October 2022. The patients were divided into training and validation groups in a 7:3 ratio. An additional external validation group included 100 lung cancer patients treated at the same institution from November 2022 to January 2024.

Inclusion criteria: 1. Patients aged 18 years or older. 2. Pathologically confirmed NSCLC [18]. 3. Availability of complete and traceable clinical examination data.

Exclusion criteria: 1. Patients transferred to another hospital. 2. Patients with multiple organ failure. 3. Patients who experienced cardiovascular or cerebrovascular accidents during the study period. 4. Patients with immune, hematologic, or liver function impairments. 5. Patients with multiple malignant tumors.

Infection diagnosis criteria

Pulmonary infection was defined by the presence of new infiltrative changes in the lungs, as shown on imaging, in combination with elevated inflammatory markers and clinical symptoms such as cough, sputum production, and fever. Diagnostic criteria followed the "Hospital Infection Diagnosis Standard" [19].

Grouping criteria

Based on the infection diagnosis criteria, 644 patients were categorized into two groups: a pulmonary infection group (238 patients) and a non-infection group (406 patients). In the training group, there were 145 patients with infections and 235 without infections. The validation group included 58 patients with infections and 106 without infections, while the external validation group consisted of 35 infection patients and 65 non-infection patients.

Feature selection and model construction

Random forest: Feature selection was performed using Recursive Feature Elimination (RFE) combined with 10-fold cross-validation. The following parameters were used: $n_{tree} = 50$, $m_{try} = \sqrt{ncol(x)}$, $importance = TRUE$. Variable importance was assessed using Mean

Nomogram for predicting lung infection risk

Decrease Accuracy and Mean Decrease Gini metrics [20].

Support vector machine (SVM): RFE combined with 10-fold cross-validation was used to select features, employing an SVM radial basis function model. Key parameters included: method = "svmRadial" trControl = trainControl (method = "cv") [21].

Lasso regression: Features were selected using the Lasso regression regularization method, with optimal λ values (λ_{min} and $\lambda_{1\text{se}}$) determined through 10-fold cross-validation. Features with non-zero coefficients were retained. Key parameters included: alpha = 1, family = "binomial" nfolds = 10 [22].

XGBoost: Feature importance was evaluated using the Gain metric from the XGBoost model. Features were ranked by Gain or cumulative contribution rates. Key parameters included: booster = "gbtree" objective = "binary" eta = 0.1 max_depth = 6 subsample = 0.7 colsample_bytree = 0.8 nrounds = 100 [23].

Baseline data collection

Patient data were retrieved from the hospital's electronic medical record system. Key variables included: gender, diabetes, hypertension, chronic obstructive pulmonary disease (COPD), chemotherapy intensity (single/combination), clinical stage (III/IV), chemotherapy cycles ($>4/\leq 4$), antibiotic use, age, body mass index (BMI), Karnofsky Performance Status (KPS), Systemic Inflammatory Index (SII), Prognostic Nutritional Index (PNI), and C-reactive protein (CRP). All data were utilized for feature screening in machine learning-based models.

Outcome measures

We compared baseline data between the pulmonary infection and non-infection groups, as well as across the training, validation, and external validation cohorts. Feature screening for pulmonary infection risk was performed using Lasso regression, XGBoost, Random Forest, and SVM. A nomogram was developed to predict pulmonary infection risk, which was validated using Receiver Operating Characteristic (ROC) curves, Precision-Recall (PR) curves, Decision Curve Analysis (DCA), and calibration curves.

Statistical analysis

Data analysis and model construction were conducted using R (versions 4.3.3 or 4.4.0). Data were imported using the data.table package and cleaned with the dplyr and tidyr packages. Continuous variables were analyzed using t-tests or rank-sum tests, while categorical variables were assessed using chi-square tests or Analysis of Variance (ANOVA). The randomForest package was used for constructing Random Forest models, and the kernlab package for SVM. Lasso regression was implemented using the glmnet package, and XGBoost modeling was performed with the xgboost package. ROC curves and AUC values were generated using the pROC package, and visualizations were created with ggplot2. Results were exported using openxlsx.

Key predictive features were identified, and machine learning models were evaluated through cross-validation, ROC curves, and feature importance rankings. Statistical significance was set at $P < 0.05$.

Results

Clinical characteristics of patients

The study compared the clinical and biological characteristics of patients in the pulmonary infection and non-infection groups. Significant differences were observed in the following variables: diabetes ($P = 0.039$), COPD ($P < 0.001$), chemotherapy intensity ($P < 0.001$), chemotherapy cycle ($P < 0.001$), antibiotic use ($P = 0.025$), age ($P < 0.001$), KPS score ($P < 0.001$), SII ($P < 0.001$), PNI ($P < 0.001$), and CRP ($P < 0.001$). No significant differences were found for gender ($P = 0.451$), hypertension ($P = 0.242$), BMI ($P = 0.218$), or clinical stage ($P = 0.286$) (**Table 1**).

Comparison of patient characteristics across groups

The baseline characteristics of the training, validation, and external validation groups were compared. No significant differences were observed in the following factors: pulmonary infection incidence ($P = 0.749$), gender ($P = 0.782$), diabetes ($P = 0.269$), hypertension ($P = 0.867$), COPD ($P = 0.180$), chemotherapy intensity ($P = 0.696$), clinical stage ($P = 0.966$),

Nomogram for predicting lung infection risk

Table 1. Comparison of baseline data between patients with pulmonary infection and the uninfected group

| Factors | Total | Pulmonary Infection Group (n = 238) | Non-infection Group (n = 406) | t/Z/ χ^2 Value | P Value |
|--------------------------|----------------------|-------------------------------------|-------------------------------|---------------------|---------|
| Gender | | | | 0.568 | 0.451 |
| Male | 342 | 131 | 211 | | |
| Female | 302 | 107 | 195 | | |
| Diabetes | | | | 4.27 | 0.039 |
| Yes | 132 | 59 | 73 | | |
| No | 512 | 179 | 333 | | |
| Hypertension | | | | 1.368 | 0.242 |
| Yes | 132 | 43 | 89 | | |
| No | 512 | 195 | 317 | | |
| COPD | | | | 18.059 | <0.001 |
| Yes | 200 | 98 | 102 | | |
| No | 444 | 140 | 304 | | |
| Chemotherapy Intensity | | | | 34.299 | <0.001 |
| Single | 297 | 74 | 223 | | |
| Combination | 347 | 164 | 183 | | |
| Clinical Stage | | | | 1.14 | 0.286 |
| III | 428 | 152 | 276 | | |
| IV | 216 | 86 | 130 | | |
| Chemotherapy Cycle | | | | 23.096 | <0.001 |
| >4 | 232 | 114 | 118 | | |
| ≤4 | 412 | 124 | 288 | | |
| Use of Antibiotics | | | | 5.005 | 0.025 |
| Yes | 312 | 129 | 183 | | |
| No | 332 | 109 | 223 | | |
| Age (years) | 24.17 ± 2.99 | 69.00 [63.00, 74.00] | 64.00 [60.00, 68.00] | 8.153 | <0.001 |
| BMI (kg/m ²) | 80.00 [70.00, 90.00] | 23.97 ± 3.18 | 24.28 ± 2.88 | -1.234 | 0.218 |
| KPS Score | 651.53 ± 202.61 | 70.00 [60.00, 80.00] | 80.00 [70.00, 100.00] | -8.564 | <0.001 |
| SII | 39.92 [36.88, 42.97] | 921.19 [807.34, 1050.80] | 608.97 [500.62, 702.66] | 23.125 | <0.001 |
| PNI | 2.25 ± 0.90 | 37.50 [33.91, 41.02] | 40.94 [38.27, 43.66] | -9.377 | <0.001 |
| CRP (g/L) | 24.17 ± 2.99 | 2.59 [2.13, 3.20] | 1.12 [0.91, 1.37] | 27.342 | <0.001 |

Note: COPD, Chronic Obstructive Pulmonary Disease; BMI, Body Mass Index; KPS, Karnofsky Performance Status; SII, Systemic Inflammatory Index; PNI, Prognostic Nutritional Index; CRP, C-Reactive Protein.

chemotherapy cycle (P = 0.803), antibiotic use (P = 0.468), age (P = 0.221), BMI (P = 0.625), KPS score (P = 0.275), SII (P = 0.668), PNI (P = 0.751), or CRP (P = 0.620). These results confirm the comparability of the groups (**Table 2**).

Evaluation and feature selection of machine learning models

In this study, we utilized Lasso regression, XGBoost, Random Forest, and SVM machine learning models to select the most predictive feature variables. The evaluation and feature selection results for each model are summarized below:

Lasso Regression: The optimal feature variables were selected based on the lambda.min and lambda.1se values, with lambda.min guiding the construction of the predictive model (**Figure 1A, 1B**). The primary feature factors included diabetes, hypertension, COPD, chemotherapy intensity, clinical stage, chemotherapy cycle, age, BMI, KPS score, SII, PNI, and CRP (**Figure 2A**).

Random Forest: The features selected by the Random Forest model included COPD, chemotherapy intensity, chemotherapy cycle, antibiotic use, age, KPS score, SII, PNI, and CRP (**Figures 1C, 2B**).

Nomogram for predicting lung infection risk

Table 2. Comparison of baseline data in training group, validation group and external validation group

| Factors | Total | Validation Group (n = 163) | Training Group (n = 381) | External Validation Group (n = 100) | F/Z/ χ^2 Value | P Value |
|--------------------------|----------------------|----------------------------|--------------------------|-------------------------------------|---------------------|---------|
| Pulmonary Infection | | | | | 0.578 | 0.749 |
| Yes | 238 | 58 | 145 | 35 | | |
| No | 406 | 106 | 235 | 65 | | |
| Gender | | | | | 0.491 | 0.782 |
| Male | 342 | 89 | 203 | 50 | | |
| Female | 302 | 75 | 177 | 50 | | |
| Diabetes | | | | | 2.625 | 0.269 |
| Yes | 132 | 28 | 86 | 18 | | |
| No | 512 | 136 | 294 | 82 | | |
| Hypertension | | | | | 0.286 | 0.867 |
| Yes | 132 | 36 | 76 | 20 | | |
| No | 512 | 128 | 304 | 80 | | |
| COPD | | | | | 3.424 | 0.180 |
| Yes | 200 | 49 | 127 | 24 | | |
| No | 444 | 115 | 253 | 76 | | |
| Chemotherapy Intensity | | | | | 0.725 | 0.696 |
| Single | 297 | 74 | 173 | 50 | | |
| Combination | 347 | 90 | 207 | 50 | | |
| Clinical Stage | | | | | 0.069 | 0.966 |
| III | 428 | 110 | 251 | 67 | | |
| IV | 216 | 54 | 129 | 33 | | |
| Chemotherapy Cycle | | | | | 0.440 | 0.803 |
| >4 | 232 | 61 | 133 | 38 | | |
| ≤4 | 412 | 103 | 247 | 62 | | |
| Use of Antibiotics | | | | | 1.519 | 0.468 |
| Yes | 312 | 82 | 177 | 53 | | |
| No | 332 | 82 | 203 | 47 | | |
| Age (years) | 24.17 ± 2.99 | 65.00 [61.00, 70.00] | 66.00 [62.00, 70.00] | 66.00 [62.00, 70.00] | 3.024 | 0.221 |
| BMI (kg/m ²) | 80.00 [70.00, 90.00] | 24.14 ± 2.87 | 24.34 ± 2.92 | 23.99 ± 3.54 | 0.471 | 0.625 |
| KPS Score | 651.53 ± 202.61 | 70.00 [70.00, 90.00] | 80.00 [70.00, 90.00] | 80.00 [70.00, 90.00] | 2.583 | 0.275 |
| SII | 39.92 [36.88, 42.97] | 696.49 [570.49, 868.98] | 690.46 [543.40, 861.96] | 704.13 [562.12, 907.97] | 0.808 | 0.668 |
| PNI | 2.25 ± 0.90 | 40.09 [37.06, 43.05] | 39.49 [36.84, 42.70] | 39.92 [36.82, 42.90] | 0.572 | 0.751 |
| CRP (g/L) | 24.17 ± 2.99 | 1.42 [1.05, 2.28] | 1.33 [1.04, 2.06] | 1.33 [0.99, 2.47] | 0.957 | 0.62 |

Note: COPD, Chronic Obstructive Pulmonary Disease; BMI, Body Mass Index; KPS, Karnofsky Performance Status; SII, Systemic Inflammatory Index; PNI, Prognostic Nutritional Index; CRP, C-Reactive Protein.

Nomogram for predicting lung infection risk

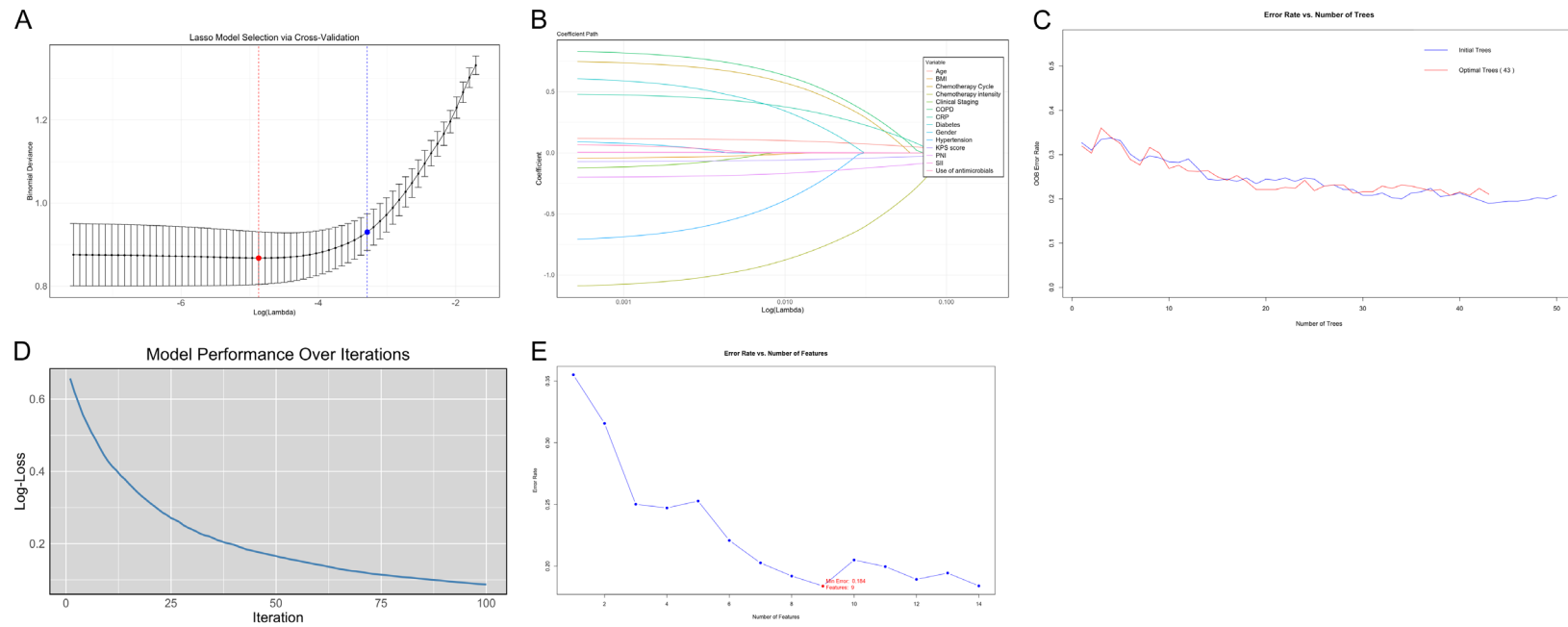


Figure 1. Evaluation and feature selection of machine learning models. A. Lasso regression path plot, with different colored lines representing the coefficient trajectories of different feature variables as the regularization parameter lambda changes. B. Lasso model selection path, with black dots representing the mean cross-validation error at each lambda value, and error bars representing the standard error. The red dashed line corresponds to lambda.min, and the blue dashed line corresponds to lambda.1se. C. XGBoost model performance plot, showing the change in log-loss value with the number of iterations. D. Random Forest error rate versus the number of trees, with the blue line representing the change in error rate with the initial number of trees and the red line representing the change in error rate with the optimized number of trees. The red dot represents the optimal number of trees (43 trees). E. RFE error rate versus the number of features, with the blue line representing the change in error rate with the number of features and the red dot representing the minimum error rate and the corresponding number of features.

Nomogram for predicting lung infection risk

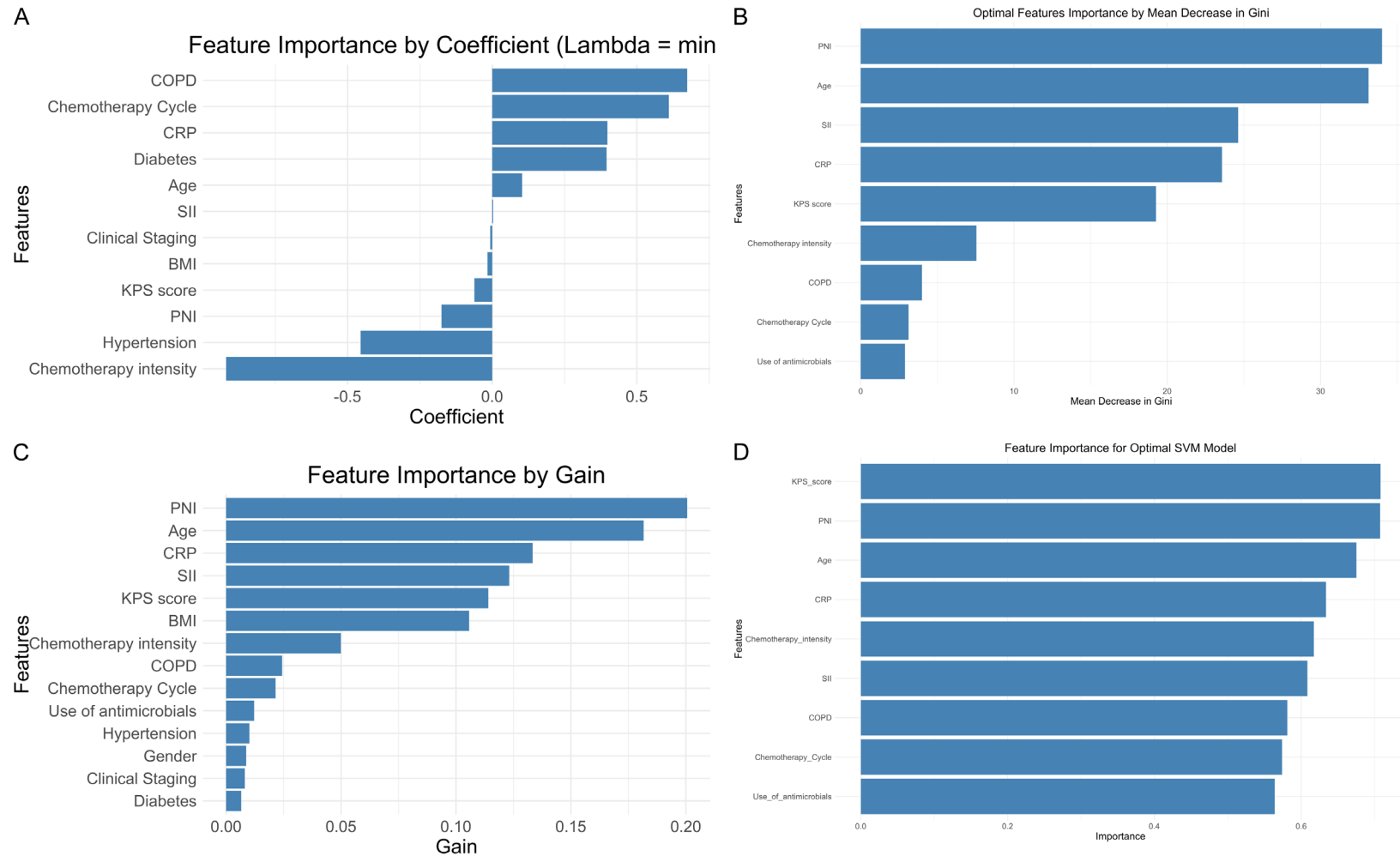


Figure 2. Feature importance display of different machine learning models. A. Lasso regression feature importance, showing the feature variables selected by the Lasso regression method at the optimal lambda value (Lambda = min) and their corresponding coefficients. Positive and negative coefficients indicate the positive and negative impact of the feature on the prediction target, respectively. B. Random Forest feature importance, showing the importance ranking of feature variables using the Random Forest method. The importance of features is measured by the Mean Decrease in Accuracy, with higher values indicating greater importance to the model. C. XGBoost feature importance, showing the importance ranking of feature variables using the XGBoost method. The importance of features is measured by Gain, indicating the average gain of the feature in all trees. Higher gain values indicate greater contribution of the feature to the model. D. SVM feature importance, showing the importance ranking of feature variables using the Support Vector Machine (SVM) method. The importance of features is measured by weight, with higher values indicating greater importance to the model.

Nomogram for predicting lung infection risk

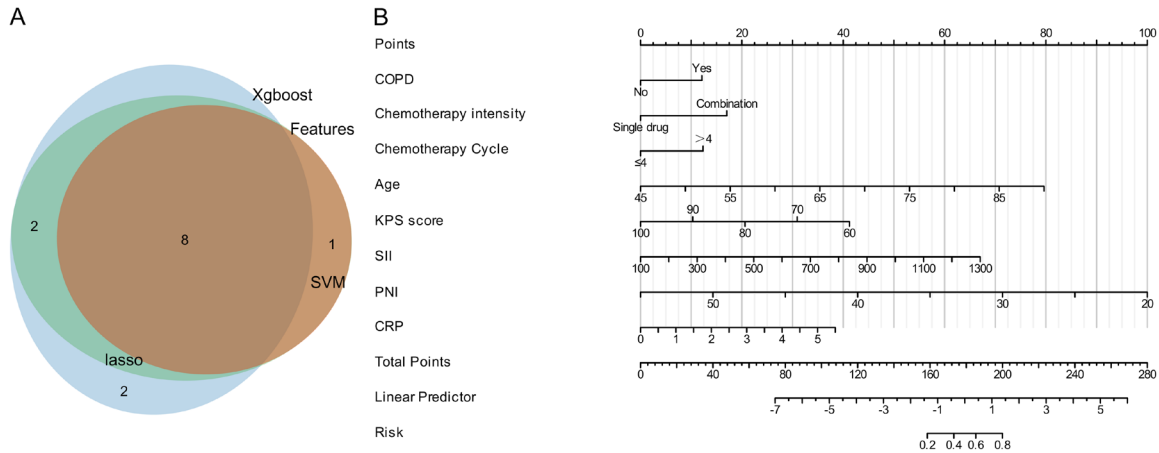


Figure 3. Feature Selection and Predictive Model Construction. A. Venn diagram showing the common feature factors selected by Lasso, SVM, XGBoost, and Random Forest, highlighting the important features identified by all four methods. B. Nomogram constructed from the 8 feature factors, used to predict the probability of pulmonary infection after radiochemotherapy. The nomogram displays the weight of each feature factor, helping to evaluate the contribution of each feature to the final prediction.

XGBoost: The XGBoost model selected the same features as Random Forest: COPD, chemotherapy intensity, chemotherapy cycle, antibiotic use, age, KPS score, SII, PNI, and CRP (**Figures 1D, 2C**).

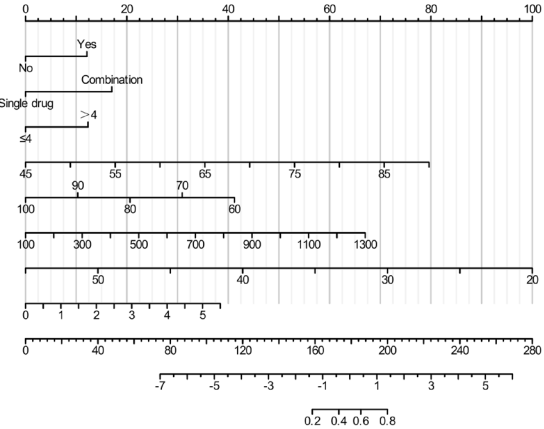
SVM: The features selected by the SVM model included KPS score, PNI, age, CRP, chemotherapy intensity, SII, COPD, chemotherapy cycle, and antibiotic use (**Figures 1E, 2D**).

Venn diagram and nomogram construction

A Venn diagram was created to display the overlap of feature variables selected by Lasso, SVM, XGBoost, and Random Forest models. The diagram identifies 8 common variables: COPD, chemotherapy intensity, chemotherapy cycle, age, KPS score, SII, PNI, and CRP (**Figure 3A**). These variables were used to construct a nomogram predicting the probability of pulmonary infection following radiochemotherapy. The nomogram illustrates the weight of each feature, helping to assess the contribution of each variable to the final prediction (**Figure 3B**).

Assessment of pulmonary infection risk factors

The assessment of pulmonary infection risk factors revealed that age, KPS score, SII, PNI, and CRP are key determinants of infection risk in lung cancer patients post-radiochemotherapy.



py. Since COPD, chemotherapy intensity, and chemotherapy cycle are binary variables, they could not be analyzed using the Restricted Cubic Spline (RCS) method. Therefore, RCS analysis was performed only for age, KPS score, SII, PNI, and CRP (**Figure 4**). Specifically:

Age exhibited a nonlinear relationship with infection risk, with a significant increase in infection risk for patients aged 70 and above (**Figure 4A**).

KPS score was negatively correlated with infection risk, with lower KPS scores indicating higher infection probabilities (**Figure 4B**).

SII was positively correlated with infection risk, with higher SII values increasing the likelihood of infection (**Figure 4C**).

PNI had a protective effect, where higher PNI values were associated with a lower infection risk (**Figure 4D**).

CRP levels demonstrated a nonlinear relationship with infection risk (**Figure 4E**). Infection risk initially increased with moderate CRP levels, plateaued briefly, and rose sharply when CRP levels exceeded a certain threshold. These findings suggest that while moderate CRP levels may have limited predictive value, significantly elevated CRP levels are strongly associated with a higher infection risk.

Nomogram for predicting lung infection risk

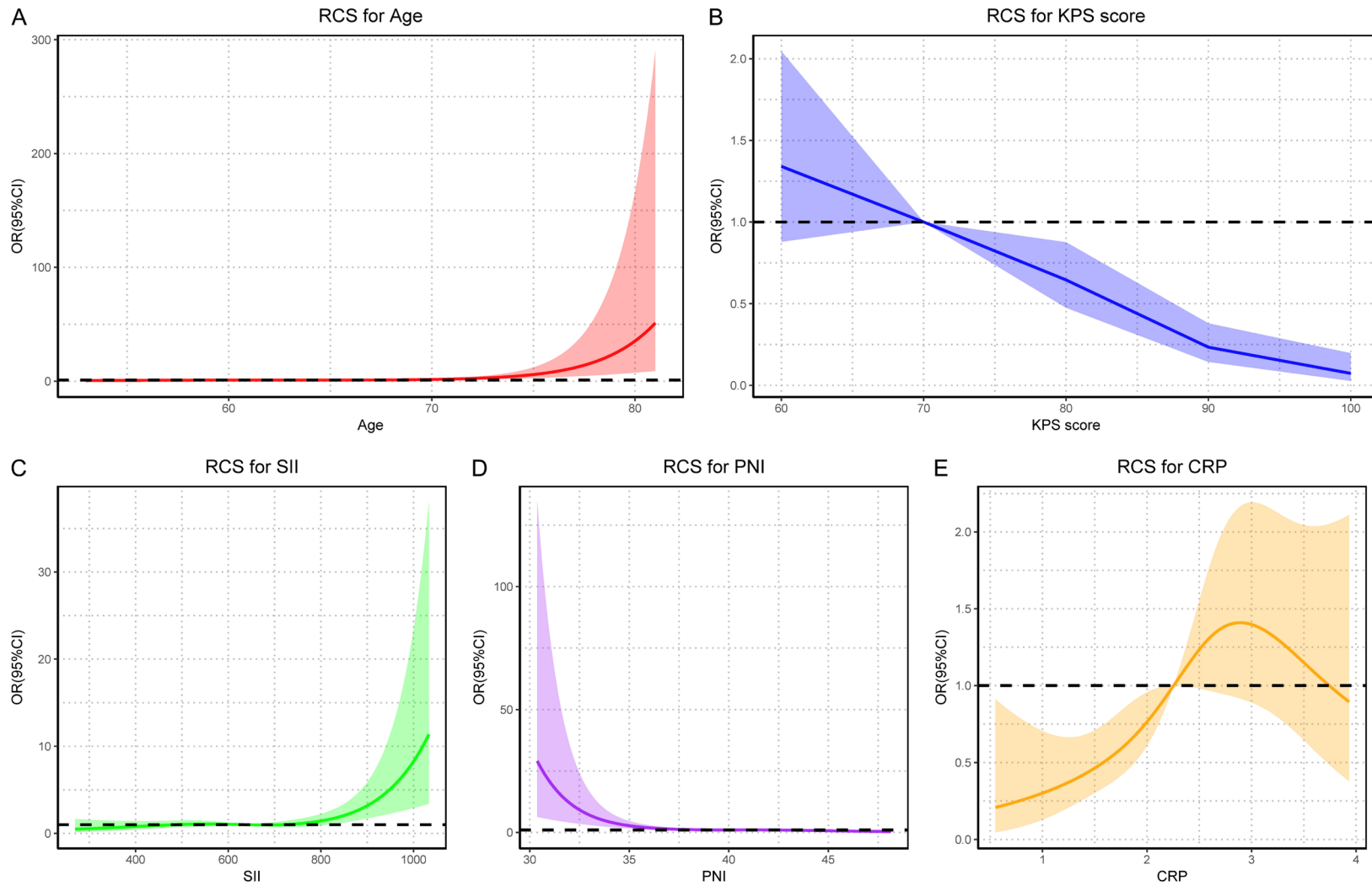


Figure 4. RCS analysis of pulmonary infection risk factors. A. RCS curve for age, showing a significant increase in pulmonary infection risk for patients aged 70 and above. B. RCS curve for KPS score, showing a significant negative correlation between lower KPS scores and higher pulmonary infection probabilities. C. RCS curve for SII, showing a positive correlation between higher SII values and increased infection risk. D. RCS curve for PNI, showing a protective effect of higher PNI values, associated with lower infection risk. E. RCS curve for CRP, emphasizing a nonlinear relationship between higher CRP levels and increased infection risk. KPS, Karnofsky Performance Status; PNI, Prognostic Nutritional Index; CRP, C-reactive protein; RCS, Restricted Cubic Spline.

Nomogram for predicting lung infection risk

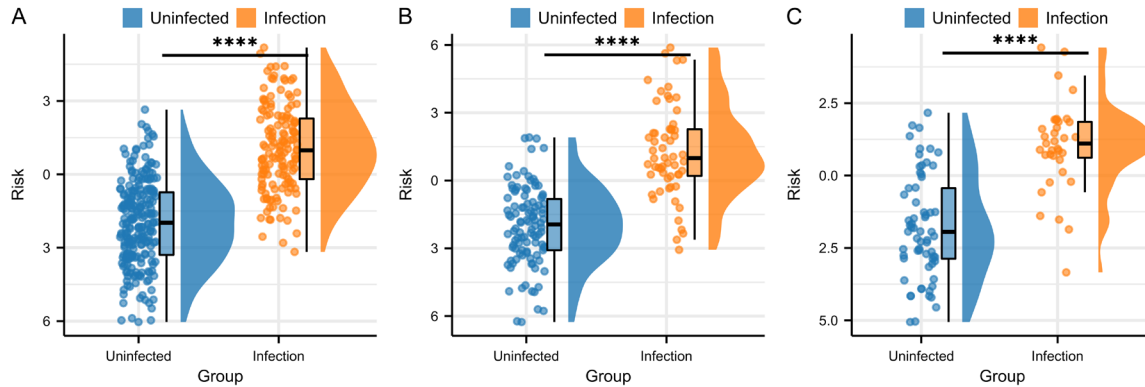


Figure 5. Comparison of risk scores between infected and non-infected patients. A. This figure shows the risk scores of infected and non-infected patients in the training set, with significant differences observed between the two groups. B. This figure shows the risk scores of infected and non-infected patients in the validation set, with significant differences observed between the two groups. C. This figure shows the risk scores of infected and non-infected patients in the external validation set, with significant differences observed between the two groups. Note: Blue: Non-infected patients; Orange: Infected patients; ****: $P < 0.0001$, indicating highly significant differences between the two groups.

Comparison of risk scores between infected and non-infected patients

Risk scores for each patient were calculated using the following formula: Risk Score = $0.8580375 + 0.809259943 * \text{COPD} + (-1.137472862) * \text{Chemotherapy intensity} + 0.823376502 * \text{Chemotherapy Cycle} + 0.118178183 * \text{Age} + (-0.068843445) * \text{KPS score} + 0.003730309 * \text{SII} + (-0.190810667) * \text{PNI} + 0.466682935 * \text{CRP}$.

Comparison of risk scores across the training, validation, and external validation groups revealed that pulmonary infection patients had significantly higher risk scores than non-infected patients ($P < 0.0001$, **Figure 5**).

ROC and PR curves for model evaluation

ROC and PR curves were constructed for the risk prediction model using the training, validation, and external validation datasets to evaluate classification ability and model discrimination.

ROC curves: Training set (AUC = 0.889, 95% CI: 0.856-0.921), validation set (AUC = 0.897, 95% CI: 0.846-0.949), external validation set (AUC = 0.875, 95% CI: 0.803-0.946) demonstrated strong classification ability and excellent discrimination (**Figure 6A-C**).

PR curves: Training set (AUC = 0.889, 95% CI: 0.856-0.921), validation set (AUC = 0.897,

95% CI: 0.846-0.949), external validation set (AUC = 0.875, 95% CI: 0.803-0.946) showed high precision and recall rates, further validating the robustness of the model (**Figure 6D-F**).

Calibration and decision curve analysis for model evaluation

We performed calibration and decision curve analyses for the risk prediction model in the training, validation, and external validation sets to assess the alignment between predicted probabilities and actual outcomes, as well as the net benefit and clinical applicability of the model at various thresholds. The calibration curves for the training set (**Figure 7A**), validation set (**Figure 7B**), and external validation set (**Figure 7C**) demonstrated good consistency between predicted probabilities and observed outcomes, indicating strong calibration performance. The decision curves for the training set (**Figure 7D**), validation set (**Figure 7E**), and external validation set (**Figure 7F**) showed a high net benefit across multiple thresholds, confirming the model's robust clinical applicability in each dataset. The highest net benefit rates were 61.84%, 64.63%, and 65.00%, respectively.

Discussion

Lung cancer remains one of the leading causes of cancer-related mortality worldwide [24]. For patients with advanced NSCLC who are in-

Nomogram for predicting lung infection risk

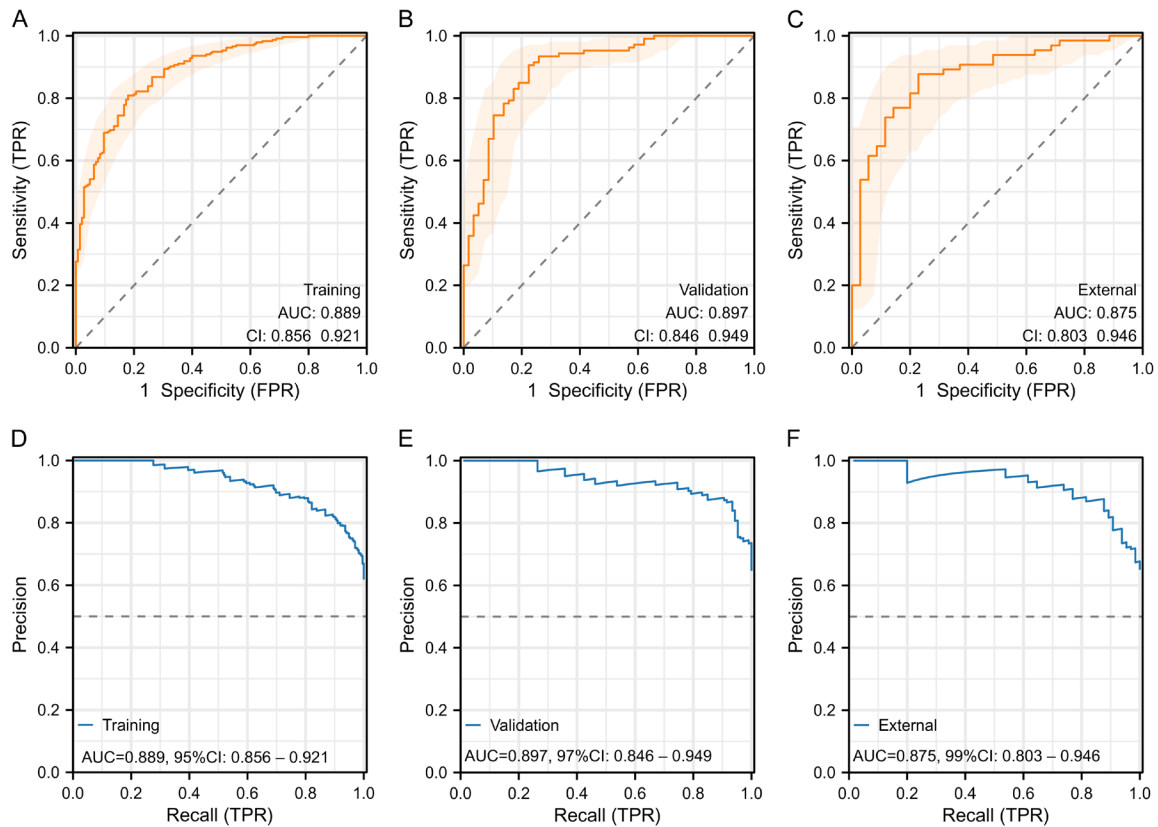


Figure 6. ROC and PR curves for model evaluation. A-C: ROC curves in the training set, validation set, and external validation set to evaluate the classification ability and discrimination of the model. D-F: PR curves in the training set, validation set, and external validation set to evaluate the precision and recall rates of the model at different thresholds. ROC, Receiver Operating Characteristic; PR, Precision-Recall.

operable or present high surgical risks, chemo-radiotherapy is a primary treatment modality [25]. While chemoradiotherapy effectively controls tumor growth, it also significantly increases the risk of pulmonary infections [19]. Therefore, accurately predicting the risk of pulmonary infections in lung cancer patients post-chemoradiotherapy is crucial for early intervention and prevention strategies. In this study, we identified eight key factors influencing the risk of pulmonary infection in these patients using Lasso regression, XGBoost, Random Forest, and SVM: COPD, chemotherapy intensity, chemotherapy cycle, age, KPS score, SII, PNI, and CRP. Based on these factors, we successfully developed a predictive model.

COPD is a major comorbidity in lung cancer patients, particularly those undergoing chemoradiotherapy [26]. COPD patients already have impaired pulmonary function, and chemoradio-

therapy further weakens the respiratory system, increasing susceptibility to infections. Chronic airway inflammation in COPD patients also provides an environment conducive to pathogen colonization and proliferation [27]. Recent research by Sun et al. [28] identified COPD as a risk factor for radiation pneumonitis in patients with esophageal squamous cell carcinoma undergoing radiotherapy. Additionally, a retrospective study showed that NSCLC patients with COPD have a significantly higher risk of severe radiation pneumonitis after radical radiotherapy compared to those without COPD [29]. These findings underscore the role of COPD in increasing the risk of pulmonary infections in lung cancer patients undergoing chemoradiotherapy.

High-intensity chemotherapy, while effective in inhibiting tumor growth, can severely damage normal cells, particularly immune cells [30]. Chemotherapy induces myelosuppres-

Nomogram for predicting lung infection risk

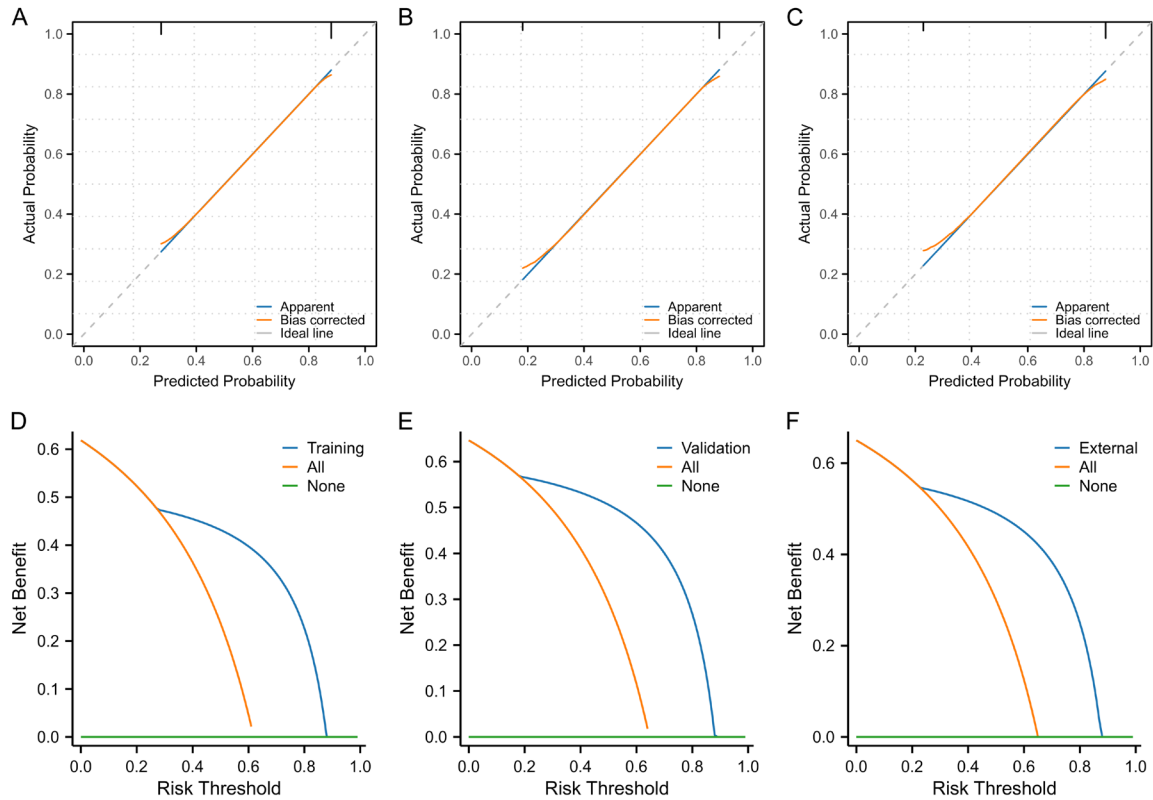


Figure 7. Calibration and DCA for model evaluation. A-C: Calibration curves in the training set, validation set, and external validation set to assess the consistency between predicted probabilities and actual outcomes. D-F: Decision curve analysis in the training set, validation set, and external validation set to evaluate the net benefit and clinical applicability of the model at different thresholds. DCA, Decision Curve Analysis.

sion, leading to significantly reduced white blood cell counts, impairing immune function and increasing infection risk [31]. The duration of chemotherapy cycles also influences infection risk. Prolonged chemotherapy cycles extend exposure to chemotherapeutic agents, further weakening immune function [32]. This prolonged exposure increases the likelihood of encountering infection sources, as the patient's immune defenses remain compromised. Strategically scheduling chemotherapy cycles can help balance tumor control and minimize infection risk. Guo et al. [33] reported that lung cancer patients undergoing combined chemotherapy and radiotherapy for more than two weeks had a 1.792-fold higher risk of developing pulmonary infections compared to those treated for less than two weeks. Ding et al. [34] found that patients receiving combined chemotherapy for ≥ 2 weeks had a 3.913-fold higher risk of postoperative pulmonary infections compared to those treated with single-agent chemotherapy, and a 4.888-fold

higher risk compared to those with chemotherapy cycles < 2 weeks. These studies highlight that high-intensity and prolonged chemotherapy cycles significantly increase the infection risk in lung cancer patients undergoing chemoradiotherapy.

Age is another critical risk factor, with patients aged 70 and above exhibiting a significantly increased risk of infections. Aging is associated with a decline in immune function, leading to slower and less effective immune responses, which in turn reduce infection resistance [35]. Additionally, older individuals often have comorbid chronic diseases, further elevating their risk of infections. The KPS score, which assesses the functional status of cancer patients, is also an important factor. A lower KPS score indicates poorer overall health and reduced functional capacity, making patients more vulnerable to infections [36]. Such patients are more likely to experience complications, including infections, during chemora-

Nomogram for predicting lung infection risk

diotherapy. Guo et al. [33] observed that lung cancer patients older than 60 years had a significantly higher risk of abdominal infections during radiotherapy. Furthermore, a meta-analysis [37] indicated that older age and COVID-19 infection are associated with an increased mortality risk in lung cancer patients.

The Systemic Immune-Inflammation Index (SII) is a comprehensive indicator of systemic inflammation, incorporating neutrophil, lymphocyte, and platelet counts [38]. Elevated SII values reflect high levels of inflammation, suggesting an activated yet dysregulated immune system, which increases the risk of infections [39]. High SII levels are significantly correlated with an elevated infection risk. A retrospective study demonstrated that elevated SII levels correlate with clinical manifestations of interstitial lung disease combined with pneumonia, with high SII serving as an independent prognostic factor for 90-day mortality [40]. The Prognostic Nutritional Index (PNI), based on serum albumin levels and lymphocyte counts, evaluates nutritional status. Higher PNI values indicate better nutritional status, which enhances immune function and infection resistance [41]. Well-nourished patients tend to recover more rapidly after chemoradiotherapy, reducing the infection risk. Ma et al. [42] identified PNI as an independent risk factor for predicting pulmonary infections after D2 radical gastrectomy in gastric cancer patients. Hazer et al. [43] found that lung cancer patients with a PNI > 50 had a postoperative infection rate of 15.5%, compared to 38.1% in those with a PNI < 50. CRP, an acute-phase protein, is a sensitive indicator of systemic inflammation. Elevated CRP levels indicate inflammation, potentially due to infection or other inflammatory diseases [44]. High CRP levels are closely associated with immune system activation and pathogen presence, suggesting an increased infection risk. Liu et al. [45] identified CRP as an independent predictor of postoperative pulmonary infections in cervical cancer patients undergoing laparoscopic surgery. Therefore, a comprehensive assessment of SII, PNI, and CRP can more accurately predict the risk of pulmonary infections in lung cancer patients post-chemoradiotherapy, enabling timely intervention and preventive measures.

Chemoradiotherapy is essential for treating advanced NSCLC, but it also increases the risk of pulmonary infections alongside its therapeutic benefits. Despite its importance, few predictive models are available for assessing pulmonary infections in lung cancer patients post-chemoradiotherapy. Developing an accurate predictive model is crucial for better evaluation and prevention of infection risk in clinical practice. In this study, we developed a Nomogram to predict pulmonary infection risk post-chemoradiotherapy based on eight key factors. The model demonstrated strong classification ability, discrimination, and stability, achieving AUC values of 0.889, 0.897, and 0.875 in the training, validation, and external validation sets, respectively. Other studies support the effectiveness of various predictive models in assessing pulmonary infection risk. Guo et al. [46] showed that artificial neural networks can predict pulmonary infection risk in lung cancer patients undergoing palliative chemotherapy, achieving an AUC of 0.897 ± 0.045 . Ding et al. [34] developed a Nomogram to predict postoperative pulmonary infections in NSCLC patients, achieving an AUC of 0.894 and an internal validation C-index of 0.900. Huang et al. [47] identified a risk model for predicting \geq grade 2 radiation pneumonitis in lung cancer patients treated with stereotactic body radiotherapy, with an AUC of 0.830, outperforming clinical or dosimetric models alone. Wang et al. [48] developed and validated a Nomogram to predict postoperative pulmonary infections in lung surgery patients, with an AUC of 0.794 in the development cohort and 0.849 in the validation cohort, demonstrating good predictive performance. Comparing these studies highlights the value of diverse methodologies and models in improving prediction accuracy. Our findings complement these studies and further underscore the importance of developing precise predictive models to assess and prevent pulmonary infection risk in clinical practice.

This study successfully developed a predictive model for pulmonary infection risk in lung cancer patients post-chemoradiotherapy based on eight characteristic factors. However, there are limitations. First, the retrospective design may introduce biases, such as incomplete patient history records or data omissions, which could

affect the accuracy and representativeness of the results. Second, the study was conducted at a single hospital, lacking multi-center data, which limits the generalizability of the findings. Future studies should adopt prospective designs, actively collect comprehensive data, and strictly control variables to minimize data collection biases. Expanding the study to include multi-center data will facilitate the acquisition of more diverse patient information, enhancing the model's generalizability and applicability in different clinical settings.

In conclusion, this study successfully developed a predictive model for pulmonary infection risk in lung cancer patients post-chemoradiotherapy, based on eight characteristic factors: COPD, chemotherapy intensity, chemotherapy cycle, age, KPS score, SII, PNI, and CRP. Using Lasso regression, Random Forest, XGBoost, and SVM, the model achieved AUCs of 0.889, 0.897, and 0.875 in the training, validation, and external validation sets, respectively, demonstrating strong classification ability and stability. This model serves as an effective tool for risk assessment in clinical practice, allowing healthcare providers to identify high-risk patients early for timely interventions and preventive measures, ultimately improving patient outcomes and quality of life. Future work should focus on further validating and optimizing the model to enhance its clinical applicability and utility.

Acknowledgements

This study was supported by the Research Projects of Biomedical Center of Hubei Cancer Hospital (No. 2022SWZX22), Wuhan 2022 Knowledge Innovation Project (No. 2022020-801010512) and China International Medical Exchange Foundation Special Fund for Young and Middle-aged Medical Research (No. Z-2014-06-2102).

Disclosure of conflict of interest

None.

Address correspondence to: Guang Han, Department of Radiation Oncology, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430079, Hubei, China. E-mail: nightingale9sci@163.com

References

- [1] He B, Zhao X, Pu Y, Sun R, Gao X and Liu W. Trends and projection of burden on lung cancer and risk factors in China from 1990 to 2060. *Thorac Cancer* 2024; 15: 1688-1704.
- [2] Nie H, Han Z, Nicholas S, Maitland E, Huang Z, Chen S, Tuo Z, Ma Y and Shi X. Costs of traditional Chinese medicine treatment for inpatients with lung cancer in China: a national study. *BMC Complement Med Ther* 2023; 23: 5.
- [3] Chen P, Liu Y, Wen Y and Zhou C. Non-small cell lung cancer in China. *Cancer Commun (Lond)* 2022; 42: 937-970.
- [4] Planchard D, Popat S, Kerr K, Novello S, Smit EF, Faivre-Finn C, Mok TS, Reck M, Van Schil PE, Hellmann MD and Peters S; ESMO Guidelines Committee. Metastatic non-small cell lung cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2018; 29 Suppl 4: iv192-iv237.
- [5] Provencio M, Nadal E, González-Larriba JL, Martínez-Martí A, Bernabé R, Bosch-Barrera J, Casal-Rubio J, Calvo V, Insa A, Ponce S, Reguart N, de Castro J, Mosquera J, Cobo M, Aguilar A, López Vivanco G, Camps C, López-Castro R, Morán T, Barneto I, Rodríguez-Abreu D, Serna-Blasco R, Benítez R, Aguado de la Rosa C, Palmero R, Hernando-Trancho F, Martín-López J, Cruz-Bermúdez A, Massuti B and Romero A. Perioperative nivolumab and chemotherapy in stage III non-small-cell lung cancer. *N Engl J Med* 2023; 389: 504-513.
- [6] Li Y, Zhang M, Yang C and Luo Y. Influencing factors of meaning in life in patients with advanced lung cancer undergoing radiochemotherapy: a cross-sectional survey. *Asia Pac J Clin Oncol* 2023; 19: 403-412.
- [7] Okumura H, Miyamoto A, Suzuki F and Takaya H. Acute hepatitis E infection during chemotherapy for lung cancer: a case report. *Chemotherapy* 2023; 68: 155-159.
- [8] Choi Y, Noh JM, Shin SH, Lee K, Um SW, Kim H, Pyo H, Ahn YC and Jeong BH. The incidence and risk factors of chronic pulmonary infection after radiotherapy in patients with lung cancer. *Cancer Res Treat* 2023; 55: 804-813.
- [9] Chen Z, Zhuang J, Liu M, Xu X, Liu Y, Yang S, Xie J, Lin N, Lai F and He F. Longitudinal analysis of quality of life in primary lung cancer patients with chlamydia pneumoniae infection: a time-to-deterioration model. *BMC Pulm Med* 2024; 24: 36.
- [10] Haug CJ and Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med* 2023; 388: 1201-1208.
- [11] Theodosiou AA and Read RC. Artificial intelligence, machine learning and deep learning:

Nomogram for predicting lung infection risk

- potential resources for the infection clinician. *J Infect* 2023; 87: 287-294.
- [12] Lyu F, Gao X, Ma M, Xie M, Shang S, Ren X, Liu M and Chen J. Crafting a personalized prognostic model for malignant prostate cancer patients using risk gene signatures discovered through TCGA-PRAD mining, machine learning, and single-cell RNA-sequencing. *Diagnostics (Basel)* 2023; 13: 1997.
- [13] Warkentin MT, Al-Sawaihey H, Lam S, Liu G, Diergaard B, Yuan JM, Wilson DO, Atkar-Khattra S, Grant B, Brhane Y, Khodayari-Moez E, Muriison KR, Tammemagi MC, Campbell KR and Hung RJ. Radiomics analysis to predict pulmonary nodule malignancy using machine learning approaches. *Thorax* 2024; 79: 307-315.
- [14] Nguyen QTN, Nguyen PA, Wang CJ, Phuc PT, Lin RK, Hung CS, Kuo NH, Cheng YW, Lin SJ, Hsieh ZY, Cheng CT, Hsu MH and Hsu JC. Machine learning approaches for predicting 5-year breast cancer survival: a multicenter study. *Cancer Sci* 2023; 114: 4063-4072.
- [15] Tang VW. Role of pathologists in nomogram development. *Pathology* 2023; 55: 1048-1049.
- [16] Lo SN, Ma J, Scolyer RA, Haydu LE, Stretch JR, Saw RPM, Nieweg OE, Shannon KF, Spillane AJ, Ch'ng S, Mann GJ, Gershenwald JE, Thompson JF and Varey AHR. Improved risk prediction calculator for sentinel node positivity in patients with melanoma: the melanoma institute Australia nomogram. *J Clin Oncol* 2020; 38: 2719-2727.
- [17] Fang C, Chen Z, Zhang J, Jin X and Yang M. Construction and evaluation of nomogram model for individualized prediction of risk of major adverse cardiovascular events during hospitalization after percutaneous coronary intervention in patients with acute ST-segment elevation myocardial infarction. *Front Cardiovasc Med* 2022; 9: 1050785.
- [18] Liu M, Zhang P, Wang S, Guo W and Guo Y. Comparison between novel online models and the AJCC 8th TNM staging system in predicting cancer-specific and overall survival of small cell lung cancer. *Front Endocrinol (Lausanne)* 2023; 14: 1132915.
- [19] Bao Q, Zhou H, Chen X, Yang Q and Zhou J. Characteristics and influencing factors of pathogenic bacteria in lung cancer chemotherapy combined with nosocomial pulmonary infection. *Zhongguo Fei Ai Za Zhi* 2019; 22: 772-778.
- [20] Thapelo TS, Mpoeleng D and Hillhouse G. Informed random forest to model associations of epidemiological priors, government policies, and public mobility. *MDM Policy Pract* 2023; 8: 23814683231218716.
- [21] Pal S, Peng Y, Aselisewine W and Barui S. A support vector machine-based cure rate model for interval censored data. *Stat Methods Med Res* 2023; 32: 2405-2422.
- [22] Nestler S and Humberg S. A lasso and a regression tree mixed-effect model with random effects for the level, the residual variance, and the autocorrelation. *Psychometrika* 2022; 87: 506-532.
- [23] Hou N, Li M, He L, Xie B, Wang L, Zhang R, Yu Y, Sun X, Pan Z and Wang K. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med* 2020; 18: 462.
- [24] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I and Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024; 74: 229-263.
- [25] Chang JY, Verma V, Li M, Zhang W, Komaki R, Lu C, Allen PK, Liao Z, Welsh J, Lin SH, Gomez D, Jeter M, O'Reilly M, Zhu RX, Zhang X, Li H, Mohan R, Heymach JV, Vaporciyan AA, Hahn S and Cox JD. Proton beam radiotherapy and concurrent chemotherapy for unresectable stage III non-small cell lung cancer: final results of a phase 2 study. *JAMA Oncol* 2017; 3: e172032.
- [26] Christenson SA, Smith BM, Bafadhel M and Putcha N. Chronic obstructive pulmonary disease. *Lancet* 2022; 399: 2227-2242.
- [27] Liang NC, Visger TV and Devereaux A. Mindfulness for those with COPD, asthma, lung cancer, and lung transplantation. *Am J Respir Crit Care Med* 2020; 202: P11-P12.
- [28] Sun L, Wang Y, Zhu L, Chen J, Chen Z, Qiu Z and Wu C. Analysis of the risk factors of radiation pneumonitis in patients after radiotherapy for esophageal squamous cell carcinoma. *Front Oncol* 2023; 13: 1198872.
- [29] Kim H, Hwang J, Kim SM, Choi J and Yang DS. Risk factor analysis of the development of severe radiation pneumonitis in patients with non-small cell lung cancer treated with curative radiotherapy, with focus on underlying pulmonary disease. *BMC Cancer* 2023; 23: 992.
- [30] Wang S, Li J, Dai J, Zhang X, Tang W, Li J, Liu Y, Wu X and Fan X. Establishment and validation of models for the risk of multi-drug resistant bacteria infection and prognosis in elderly patients with pulmonary infection: a multicenter retrospective study. *Infect Drug Resist* 2023; 16: 6549-6566.
- [31] Tong Y, Wen J, Yang T, Li H, Wei S, Jing M, Wang M, Zou W and Zhao Y. Clinical efficacy and safety of tanreqing injection combined with antibiotics versus antibiotics alone in the treatment of pulmonary infection patients after chemotherapy with lung cancer: a systematic

Nomogram for predicting lung infection risk

- review and meta-analysis. *Phytother Res* 2021; 35: 122-137.
- [32] Al-Mozaini MA, Islam M, Noman ASM, Karim AR, Farhat WA, Yeger H and Islam SS. Decline in respiratory functions in hospitalized SARS-CoV-2 infected cancer patients following cytotoxic chemotherapy-an additional risk for post-chemotherapy complications. *Front Med (Lausanne)* 2022; 9: 835098.
- [33] Guo L, Dong Y, Qi Y, Tao N, Song H, Shao L, Cai Y, Xu L and Wei S. Analysis of risk factors for pulmonary infections during radiotherapy in lung cancer patients. *Altern Ther Health Med* 2024; AT10312.
- [34] Ding Z, Wang X, Jiang S and Liu J. Risk factors for postoperative pulmonary infection in patients with non-small cell lung cancer: analysis based on regression models and construction of a nomogram prediction model. *Am J Transl Res* 2023; 15: 3375-3384.
- [35] Martinez L, Seddon JA, Horsburgh CR, Lange C and Mandalakas AM; TB Contact Studies Consortium. Effectiveness of preventive treatment among different age groups and mycobacterium tuberculosis infection status: a systematic review and individual-participant data meta-analysis of contact tracing studies. *Lancet Respir Med* 2024; 12: 633-641.
- [36] Choudhary NS, Sonavane A, Saraf N, Saigal S, Rastogi A, Bhargui P, Thiagrajan S, Yadav SK, Saha S and Soin AS. Poor performance status predicts mortality after living donor liver transplantation. *J Clin Exp Hepatol* 2020; 10: 37-42.
- [37] Wu M, Liu S, Wang C, Wu Y and Liu J. Risk factors for mortality among lung cancer patients with covid-19 infection: a systematic review and meta-analysis. *PLoS One* 2023; 18: e0291178.
- [38] Ozer Balin S, Ozcan EC and Uğur K. A new inflammatory marker of clinical and diagnostic importance in diabetic foot infection: systemic immune-inflammation index. *Int J Low Extrem Wounds* 2022; 15347346221130817.
- [39] Kocaaslan R, Dilli D and Çitli R. Diagnostic value of the systemic immune-inflammation index in newborns with urinary tract infection. *Am J Perinatol* 2024; 41: e719-e727.
- [40] Bai W, Wang Y and Li F. Effect of novel inflammatory biomarkers on adverse outcomes in patients with interstitial lung disease and pneumonia: a multicenter retrospective cohort study. *Comb Chem High Throughput Screen* 2024; [Epub ahead of print].
- [41] Nergiz S and Ozturk U. The effect of prognostic nutritional index on infection in acute ischemic stroke patients. *Medicina (Kaunas)* 2023; 59: 679.
- [42] Ma X, Lu X, Jiang X, Wang J, Wang T and Zhang L. A nomogram combining prognostic nutritional index and platelet lymphocyte ratio predicts postoperative pulmonary infection following D2 radical gastrectomy for gastric cancer. *Nutr Hosp* 2024; 41: 602-611.
- [43] Hazer S, Gülhan SŞE, Solak N, Yenibertiz D, Akıllı MS, Sayilir Guven E and Bıçakçioğlu P. The effect of prognostic nutritional index in postoperative infection following lobectomy in non-small cell lung cancer patients. *Cureus* 2023; 15: e37611.
- [44] Zheng S and Zhang W. Predictive values of sTREM-1, PCT and CRP for multiple trauma-induced acute respiratory distress syndrome complicated with pulmonary infection. *Clin Lab* 2022; 68: 1-13.
- [45] Liu Y, Tian L, You J and Li Y. The predictive value of postoperative C-reactive protein (CRP), procalcitonin (PCT) and triggering receptor expressed on myeloid cells 1 (TREM-1) for the early detection of pulmonary infection following laparoscopic general anesthesia for cervical cancer treatment. *Ann Palliat Med* 2021; 10: 4502-4508.
- [46] Guo W, Gao G, Dai J and Sun Q. Prediction of lung infection during palliative chemotherapy of lung cancer based on artificial neural network. *Comput Math Methods Med* 2022; 2022: 4312117.
- [47] Huang BT, Lin PX, Wang Y and Luo LM. Developing a prediction model for radiation pneumonitis in lung cancer patients treated with stereotactic body radiation therapy combined with clinical, dosimetric factors, and laboratory biomarkers. *Clin Lung Cancer* 2023; 24: e323-e331, e322.
- [48] Wang JY, Pang QY, Yang YJ, Feng YM, Xiang YY, An R and Liu HL. Development and validation of a nomogram for predicting postoperative pulmonary infection in patients undergoing lung surgery. *J Cardiothorac Vasc Anesth* 2022; 36: 4393-4402.