Big Data and Information Analytics

https://www.aimspress.com/journal/BDIA

BDIA, 9: 136–151.
DOI:10.3934/bdia.2025007
Received: 22 September 2025
Revised: 25 October 2025
Accepted: 05 November 2025
Published: 10 November 2025

Research article

Comparison of proportional hazards model and random survival forest in lung cancer survival prediction

Ceshu Zheng and Zhongqiang Liu*

School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo 454003, China

* **Correspondence:** Email: zhongqiang@hpu.edu.cn.

Abstract: This study develops and compares a traditional Cox proportional hazards nomogram with a machine learning-based random survival forest (RSF) model to predict 1-, 3-, and 5-year overall survival in lung cancer patients using data from the SEER database (N = 8695). Variable selection is conducted through univariate screening ($p < 0.05$), multivariate Cox regression, and RSF variable importance (VIMP) analysis, identifying key prognostic factors, such as TNM stage, tumor size, chemotherapy status, and patient age. Compared with the Cox proportional hazards model, the RSF model has a slightly lower concordance index (C-index), but its time-dependent area under the curve (AUC) values are all higher. This indicates that the RSF model is more accurate in prediction at specific time points, confirming its clinical utility. Both models also exhibit good calibration. These interpretable prognostic tools provide valuable support for clinical decision making, enabling improved risk stratification and personalized survival estimation. The findings underscore the potential of machine learning approaches to enhance predictive accuracy in oncology.

Keywords: lung cancer; Cox proportional hazards model; random survival forest model; survival prediction

1. Introduction

Lung cancer are among the most common and lethal malignancies worldwide, with both incidence and mortality rates ranking highest among all cancer types. According to World Health Organization estimates, approximately 2 million new cases of lung cancer are diagnosed annually, resulting in about 1.76 million deaths. This disease imposes substantial psychological and economic burdens on patients and their families [1]. Despite considerable advances in diagnostic and therapeutic techniques in recent years, the overall prognosis for lung cancer remains poor.

The high mortality rate is largely due to the lack of noticeable early symptoms, which leads to the

majority of patients being diagnosed at advanced stages. Late-stage diagnosis is associated with limited treatment options and unfavorable outcomes. Therefore, accurate prediction of patient prognosis is critical for developing personalized treatment strategies and improving survival rates.

The survival outcomes of cancer patients are influenced by a variety of factors, including clinical and pathological characteristics as well as treatment strategies. In recent years, the Cox proportional hazards model has been widely employed in survival analysis due to its ability to integrate multiple predictors and provide individualized survival estimates. Nomograms developed based on this model have demonstrated high predictive accuracy and are increasingly utilized in clinical decision support. For instance, Zhu et al. constructed and validated a nomogram for predicting survival in patients with second primary small cell lung cancer using univariate and multivariate Cox regression analyses, incorporating variables, such as disease stage, age, chemotherapy, radiotherapy, and surgical history [2]. Similarly, Alexander et al. developed the lung cancer prognostic index for non-small cell lung cancer by integrating factors, including age, sex, and disease stage, through Cox regression modeling to predict overall survival [3].

Nevertheless, the impact of socio-economic factors, such as marital status, household income, and urban versus rural residence, on lung cancer prognosis remains insufficiently investigated.

Meanwhile, the rapid advancement of machine learning has drawn increasing attention to its applications in survival analysis. As a non-parametric tree-based method, the random survival forest (RSF) model is capable of capturing complex non-linear relationships and feature interactions, while demonstrating strong adaptability to high-dimensional data [4]. Unlike the Cox model, RSF does not require the proportional hazards assumption, thereby offering greater flexibility and robustness in handling complex datasets.

For example, Churpek et al. compared the predictive performance of RSF, logistic regression, the modified early warning score, and conventional regression models in forecasting clinical deterioration among heart disease patients. Their results showed that although all models provided predictive value, the RSF model achieved the highest performance [5]. Thus, the incorporation of RSF into survival prediction for lung cancer may provide multidimensional support for individualized prognostic assessment.

This study systematically compares the Cox proportional hazards model with the random survival forest (RSF) model for predicting overall survival in lung cancer patients using SEER database. Two key innovations distinguish this research: first, the inclusion of socioeconomic variables alongside traditional clinicopathological factors to develop a more comprehensive prognostic model; second, systematic hyperparameter tuning of RSF through grid search to enhance its capability in capturing complex nonlinear relationships. This robust methodological comparison provides a more reliable foundation for prognostic assessment in lung cancer.

2. Methodology for prognostic modeling in lung cancer

A total of 12,072 patients diagnosed with lung cancer between 2010 and 2018 were retrieved from the surveillance, epidemiology, and end results (SEER) database using SEER*Stat software (version 8.4.3). This study utilized de-identified data from the SEER database and was therefore exempt from institutional review board review and the requirement for informed consent.

To ensure the robustness and generalizability of our findings, all eligible cases from the database

were included. Consistent with established methodologies in the field [6, 7] the inclusion criteria were as follows: (1) age at diagnosis between 0 and 100 years, (2) availability of complete follow-up information, and (3) presence of a single primary lung cancer. Patients with a history of additional primary malignancies that could potentially confound survival prediction were excluded. Exclusion criteria comprised: (1) missing values for any variable included in the subsequent analysis and (2) a recorded survival time of less than one month. The overall study selection process is delineated in Figure 1.

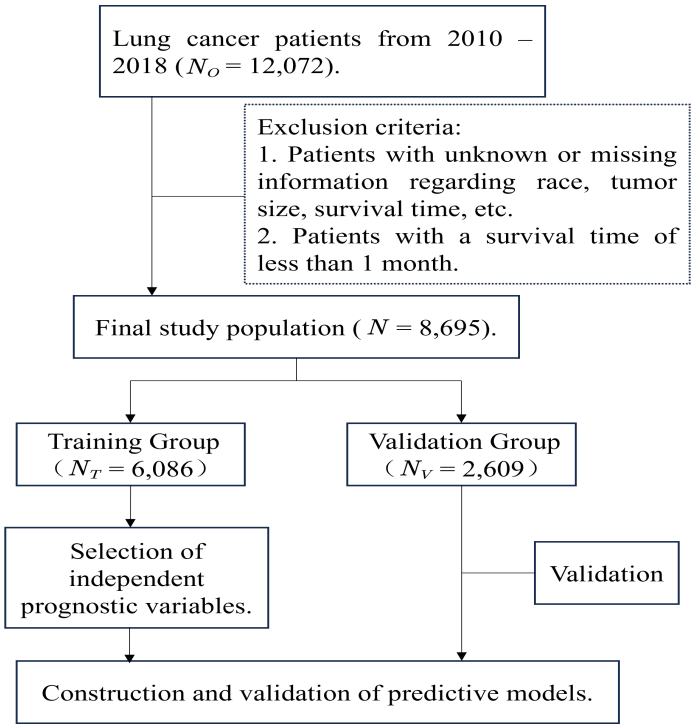


Figure 1. The overall study selection process.

Based on previously established literature and clinical relevance, demographic and clinical characteristics of the patients were extracted from the SEER database. Demographic variables included age at diagnosis, sex, race, marital status, total income, urban versus rural residence, and survival time (in months). Clinical characteristics included tumor size, tumor grade, AJCC stage (determined based on the T [tumor], N [lymph node], and M [distant metastasis] categories of the primary tumor), chemotherapy status, and tumor laterality.

Based on the distribution of the data, the original variable categorizations from the SEER database were modified to ensure a sufficient sample size in each group and to make the data more amenable to statistical analysis. Age was categorized into five groups: <50, 50–59, 60–69, 70–79, and ≥80 years. This recategorization scheme retained the essential information of the variables while enhancing the validity of subsequent statistical comparisons. Laterality was classified into three categories based on tumor location: left-sided, right-sided, and bilateral. Tumor grade was grouped into four categories: grades I through IV. Total income was divided into six intervals: <40,000, 40,000–59,999, 60,000–79,999, …, and ≥120,000. The primary endpoint of the study was overall survival (OS), defined as the duration from the date of diagnosis to death due to cancer [8].

2.1. The Cox proportional hazards model and the RSF model

Univariate Cox regression analysis was initially conducted to identify variables associated with prognosis. Variables with a p-value < 0.05 were subsequently incorporated into a multivariate Cox proportional hazards model to determine independent prognostic factors in patients with lung and bronchial cancer. These independent predictors were used to construct a nomogram for individualized outcome prediction.

The RSF is an ensemble tree-based method for the analysis of right-censored survival data. Its algorithm operates through the following key steps: (1) **bootstrap sampling**: draw multiple bootstrap samples with replacement from the original training set to build individual survival trees; (2) **random feature selection**: at each node, randomly select `mtry` candidate predictor variables and choose the optimal split point based on the log-rank statistic; (3) **tree growth control**: grow each tree fully under the constraint that terminal nodes must contain at least `nodesize` unique events to control overfitting; and (4) **ensemble prediction**: for a new patient, the ensemble cumulative hazard function (CHF) is obtained by averaging the CHFs across all trees, from which the survival probability is derived.

The survival of lung cancer patients was modeled and predicted using the randomForestSRC package in R. Following a common practice similar to that used for Cox proportional hazards modeling, the entire dataset was randomly partitioned into a training group (70%) and a validation group (30%).

2.2. Survival analysis and model evaluation

All statistical analyses in this study were conducted using R software (version 4.4.0), available for download at https://www.r-project.org/. Continuous variables are presented as mean with standard deviation or median with interquartile range, as appropriate. The mean represents the arithmetic average and serves as a measure of central tendency. The 95% confidence interval (95% CI) provides a range within which the true population parameter is expected to lie with 95% confidence. The hazard ratio (HR), derived from Cox regression models, was used to compare the risk of the outcome event between groups during the follow-up period.

The Kaplan-Meier curve is a fundamental tool in survival analysis, used to visualize the probability of an event occurring over time [9, 10]. The survival probability changes only at each time point where an event is observed. The survival function is calculated as follows:

$$S(t) = S(t-1) \times (1 - \frac{d_t}{N_t}),$$

where $S(t)$ represents the survival probability at time t, $S(t-1)$ denotes the survival probability immediately before time t, d_t is the number of events (e.g., deaths) occurring at time t, and N_t refers to the number of subjects at risk (i.e., still under observation and event-free) just prior to time t.

The discriminative ability of the model was evaluated using the Harrell's concordance index (C-index), with higher values indicating better performance in distinguishing between patients with different survival outcomes [11]. Additionally, the predictive accuracy of the model was assessed using time-dependent receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC). Calibration curves were generated using bootstrap resampling with 1000 iterations to evaluate the agreement between predicted and observed survival probabilities.

Similar to variable screening in univariate and multivariate Cox regression, the variable importance (VIMP) measure in the RSF model was computed to quantify the contribution of each predictor. A higher VIMP value indicates a stronger influence on the prediction of survival probability.

3. Development and validation of the Cox proportional hazards regression model

Data from 12,072 patients diagnosed between 2010 and 2018 were retrieved from the SEER database. After applying inclusion and exclusion criteria, 8695 patients were included in the final analysis. These patients were randomly allocated into training and validation groups at a 7 : 3 ratio. The baseline characteristics of the study population are summarized in Table 1. Both chi-square and Kolmogorov-Smirnov (K-S) tests yielded p-values greater than 0.05, indicating no statistically significant differences between the training and validation groups in terms of baseline characteristics.

As shown in Table 1, the study cohort predominantly consisted of elderly patients. The majority were Caucasian (77.96%), followed by African American (6.87%) and other racial groups (15.17%). Most patients (54.89%) presented with Grade III tumors, while Grade I, II, and IV tumors accounted for 8.41%, 30.64%, and 6.06% of cases, respectively. Chemotherapy was administered to 62.69% of the patients. Tumors were primarily unilateral, with a small proportion (0.53%) exhibiting bilateral involvement.

The resulting curve is a step function that begins at 100% and decreases at each event time. Figure 2 illustrates the Kaplan-Meier survival analysis for all lung cancer patients in the study cohort.

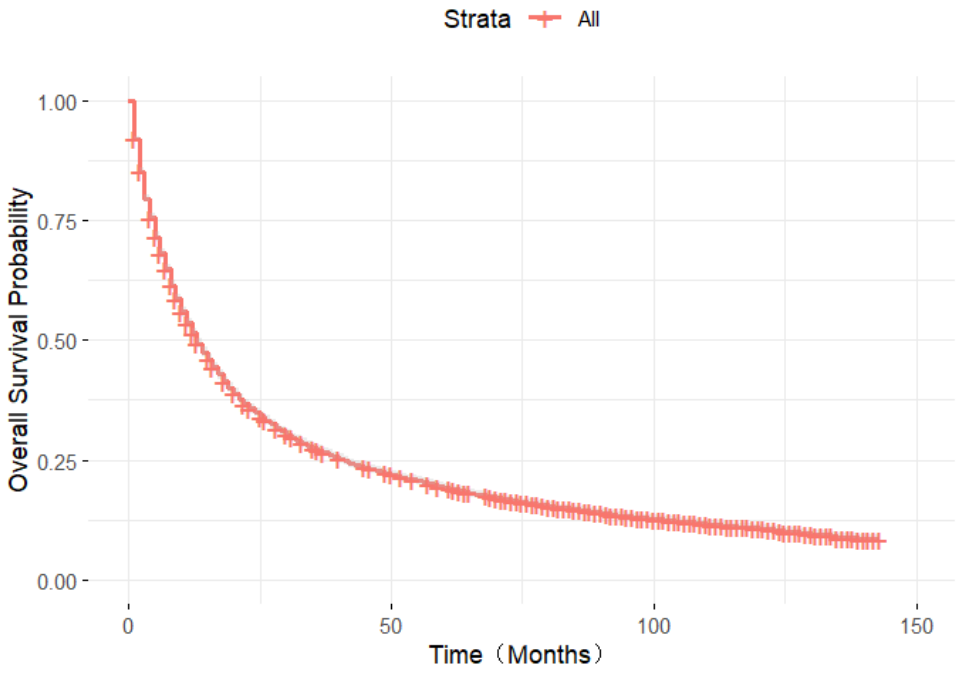


Figure 2. Kaplan-Meier survival curve for all lung cancer patients.

Table 1. Clinical and pathological characteristics of the 8695 lung cancer patients.

Variable	Total (N = 8695)	Training (N_T = 6086)	Validation (N_V = 2609)	P-value
Age (n[%])				0.3999
<50	325 (3.74)	240 (3.94)	85 (3.26)	
50–59	1283 (14.76)	882 (14.49)	401 (15.37)	
60–69	2754 (31.67)	1918 (31.51)	836 (32.04)	
70–79	2974 (34.20)	2100 (34.51)	874 (33.50)	
≥80	1359 (15.63)	946 (15.54)	413 (15.83)	
Sex (n[%])				0.1443
Female	3267 (37.57)	2256 (37.07)	1011 (38.75)	
Male	5428 (62.43)	3830 (62.93)	1598 (61.25)	
Race (n[%])				0.9437
Black	597 (6.87)	421 (6.92)	176 (6.75)	
White	6779 (77.96)	4745 (77.97)	2034 (77.96)	
Other	1319 (15.17)	920 (15.12)	399 (15.29)	
T stage (n[%])				0.1664
T0	26 (0.30)	22 (0.36)	4 (0.15)	
T3	4583 (52.71)	3201 (52.60)	1382 (52.97)	
T4	3597 (41.37)	2535 (41.65)	1062 (40.71)	
Tx	489 (5.62)	328 (5.39)	161 (6.17)	
N stage (n[%])				0.3153
N0	3009 (34.61)	2127 (34.95)	882 (33.81)	
N1	973 (11.19)	694 (11.40)	279 (10.69)	
N2	3163 (36.38)	2182 (35.85)	981 (37.60)	
N3	1192 (13.71)	823 (13.52)	369 (14.14)	
Nx	358 (4.12)	260 (4.27)	98 (3.76)	
M stage (n[%])				0.5171
M0	4387 (50.45)	3095 (50.85)	1292 (49.52)	
M1a	1338 (15.39)	927 (15.23)	411 (15.75)	
M1b	2970 (34.16)	2064 (33.91)	906 (34.73)	
Grade (n[%])				0.1539
Well (I)	731 (8.41)	514 (8.45)	217 (8.32)	
Moderately (II)	2664 (30.64)	1887 (31.01)	777 (29.78)	
Poorly (III)	4773 (54.89)	3338 (54.85)	1435 (55.00)	
Undifferentiated (IV)	527 (6.06)	347 (5.70)	180 (6.90)	
Chemotherapy (n[%])				0.8889
No/Unknown	3244 (37.31)	2274 (37.36)	970 (37.18)	
Yes	5451 (62.69)	3812 (62.64)	1639 (62.82)	
Laterality (n[%])				0.8919
Left	3620 (41.63)	2525 (41.49)	1095 (41.97)	
Right	5029 (57.84)	3528 (57.97)	1501 (57.53)	
Both	46 (0.53)	33 (0.54)	13 (0.50)	
Marital (n[%])				0.4864
Married	8203 (94.34)	5749 (94.46)	2454 (94.06)	
Unknown	492 (5.66)	337 (5.54)	155 (5.94)	
Other	25 (0.29)	17 (0.28)	8 (0.31)	
Income (n[%])				0.2933
< 40000	775 (8.91)	558 (9.17)	217 (8.32)	
40000–59999	4519 (51.97)	3182 (52.28)	1337 (51.25)	
60000–79999	2299 (26.44)	1567 (25.75)	732 (28.06)	
80000–99999	893 (10.27)	629 (10.34)	264 (10.12)	
100000–119999	184 (2.12)	133 (2.19)	51 (1.95)	
≥120000	1057 (12.16)	738 (12.12)	319 (12.23)	
Residence (n[%])				0.9236
Rural	7638 (87.84)	5348 (87.88)	2290 (87.77)	
Urban	1057 (12.16)	738 (12.12)	319 (12.23)	
Tumor-size (Mean ± SD)	62.73 ± 79.42	63.00 ± 80.18	62.09 ± 77.60	0.8508

A univariate Cox proportional hazards model was used to evaluate the association between a single predictor variable (e.g., age, sex, or chemotherapy) and time-to-event outcomes, such as overall survival. This model computes a hazard ratio for the variable, which reflects the instantaneous risk of the event occurring in one group relative to a reference group, without accounting for other factors. Univariate analysis serves as an initial screening tool to identify potential predictors worthy of further investigation.

In contrast, the multivariate Cox proportional hazards model examines the relationship between multiple predictors and the time-to-event outcome simultaneously. It aims to estimate the independent effect of each variable on the hazard after adjusting for potential confounders included in the model. For instance, this approach can isolate the effect of a new drug on survival while controlling for covariates, such as age, cancer stage, and sex.

Table 2. Univariate and multivariate Cox regression analyses for lung cancer patients.

Variable	Univariate analysis		Multivariate analysis	
	HR (95% CI)	P-value	HR (95% CI)	P-value
Age				
<50	Reference	< 0.001	Reference	–
50–59	1.248 (1.059–1.470)	–	1.241 (1.053–1.463)	0.010
60–69	1.341 (1.149–1.565)	0.008	1.527 (1.307–1.785)	< 0.001
70–79	1.612 (1.383–1.879)	< 0.001	1.937 (1.657–2.264)	< 0.001
≥80	2.191 (1.866–2.573)	< 0.001	2.562 (2.172–3.022)	< 0.001
Sex				
Female	Reference	< 0.001	Reference	–
Male	1.335 (1.261–1.412)	–	1.247 (1.177–1.320)	< 0.001
Race				
White	Reference	0.002	Reference	–
Black	1.111 (0.999–1.234)	–	1.130 (1.015–1.257)	0.025
Other	0.899 (0.832–0.970)	0.050	0.834 (0.770–0.902)	< 0.001
T stage				
T0	Reference	< 0.001	Reference	–
T3	1.109 (0.698–1.763)	–	2.032 (1.273–3.245)	0.003
T4	1.477 (0.929–2.349)	0.662	2.219 (1.390–3.543)	< 0.001
Tx	1.986 (1.235–3.194)	0.099	2.318 (1.437–3.738)	< 0.001
N stage				
N0	Reference	< 0.001	Reference	–
N1	1.307 (1.189–1.436)	–	1.389 (1.262–1.530)	<0.001
N2	2.011 (1.883–2.148)	< 0.001	1.832 (1.703–1.970)	< 0.001
N3	2.354 (2.159–2.566)	< 0.001	1.878 (1.708–2.065)	< 0.001
Nx	2.272 (1.987–2.598)	< 0.001	1.725 (1.499–1.985)	< 0.001
M stage				
M0	Reference	< 0.001	Reference	–
M1a	1.697 (1.569–1.836)	–	1.614 (1.487–1.752)	< 0.001
M1b	2.748 (2.585–2.921)	<0.001	2.725 (2.545–2.919)	< 0.001
Grade				
Well (I)	Reference	<0.001	Reference	-
Moderately (II)	1.577 (1.407–1.767)	–	1.371 (1.222–1.538)	<0.001
Poorly (III)	2.081 (1.866–2.321)	< 0.001	1.680 (1.502–1.880)	< 0.001
Undifferentiated (IV)	2.643 (2.274–3.072)	< 0.001	1.951 (1.672–2.277)	< 0.001
Chemotherapy				
Yes	Reference	< 0.001	Reference	–
No/Unknown	1.127 (1.065–1.192)	–	1.611 (1.513–1.716)	< 0.001
Tumor-size	1.001 (1.001–1.001)	< 0.001	1.001 (1.001–1.001)	< 0.001

In this study, both univariate and multivariate Cox regression models were employed to analyze the associations between predictor variables and survival outcomes. The results of these analyses are summarized in Table 2.

A nomogram is a graphical calculating device designed to simplify statistical predictive models by providing a visual representation of the relationships between multiple predictors and a clinical outcome. It enables clinicians to estimate an individual patient's probability of a specific event, such as disease risk or survival, by summing points corresponding to different variable values. The primary purpose of a nomogram is to support rapid and user-friendly application of complex models for personalized risk assessment and clinical decision-making at the point of care.

Variables that demonstrated statistical significance in both univariate and multivariate Cox regression analyses were identified as independent prognostic factors. These included age, gender, race, TNM stage, tumor grade, and five additional variables. Using these predictors, a prognostic nomogram was constructed from the training group to estimate the probabilities of 1-, 3-, and 5-year overall survival in patients with lung cancer. The final prognostic nomogram is presented in Figure 3.

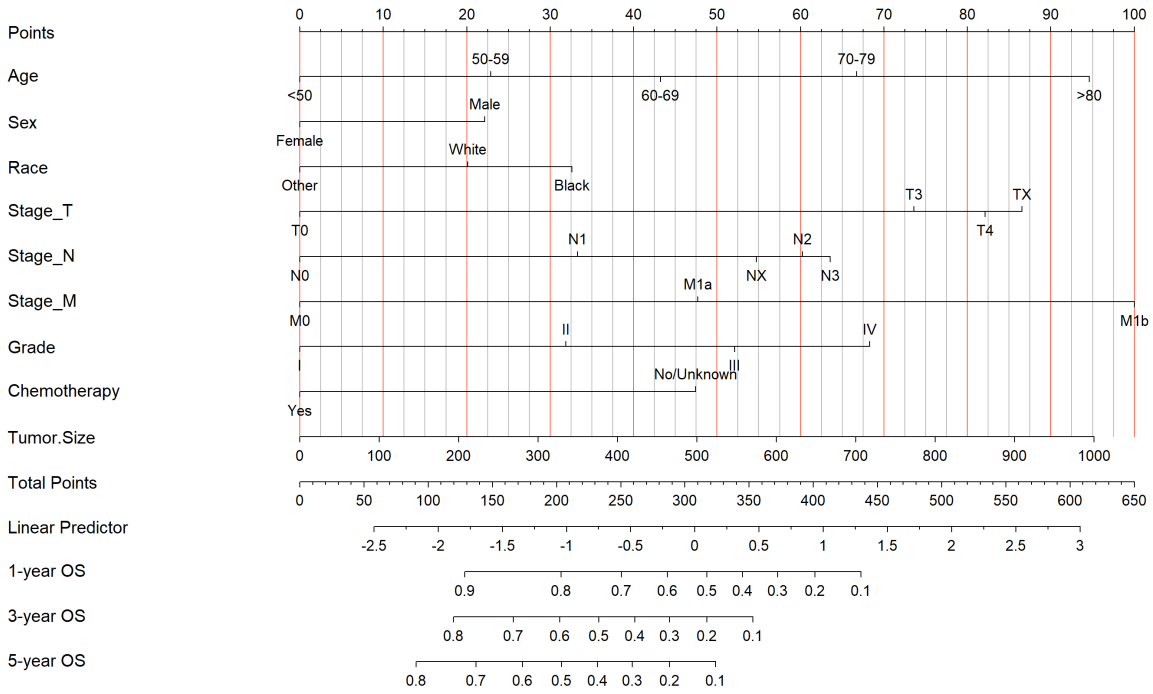


Figure 3. Nomogram for predicting the 1-, 3-, and 5-year overall survival rates of lung cancer patients.

Using this prediction model to calculate the total score, we found that the median score for all patients was 16. Patients with a total score ≥16 were classified as high-risk, while those with a score <16 were classified as low-risk [12] In both the training and validation groups, Kaplan-Meier survival curves revealed a statistically significant disparity in survival outcomes between the two groups, as assessed by the log-rank test ($p < 0.0001$). This result indicates that patients in the

high-risk group had a significantly poorer prognosis compared to their low-risk counterparts. Figure 4 illustrates the Kaplan-Meier survival curves for the two risk strata across both groups.

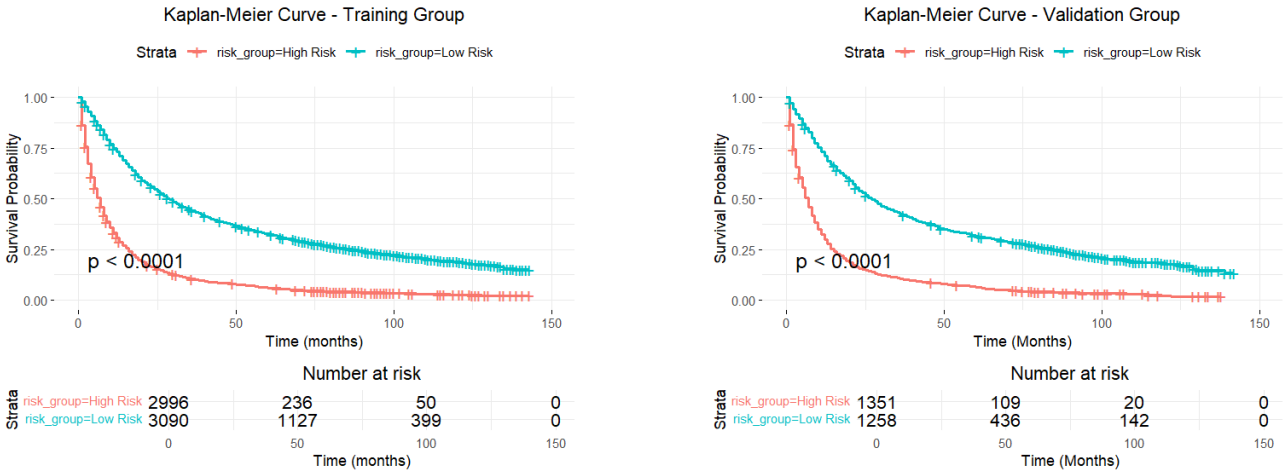


Figure 4. Kaplan-Meier survival curves for the two mortality risk subgroups in the training and validation groups.

The C-index was 0.707 in the training group and 0.702 in the validation group, indicating adequate discriminative ability of the prediction model. The model's predictive performance was further evaluated using time-dependent ROC analysis. The area under the ROC curve (AUC) values for the training group were 0.779 at 1 year, 0.784 at 3 years, and 0.792 at 5 years. Corresponding AUC values for the validation group were 0.767, 0.779, and 0.790, respectively, consistently demonstrating the strong discriminatory power of the nomogram.

Furthermore, 5-fold cross-validation yielded a C-index of 0.707, with AUC values of 0.777, 0.782, and 0.790 at 1, 3, and 5 years, respectively. Detailed results of these analyses are presented in Figures 5 and 6.

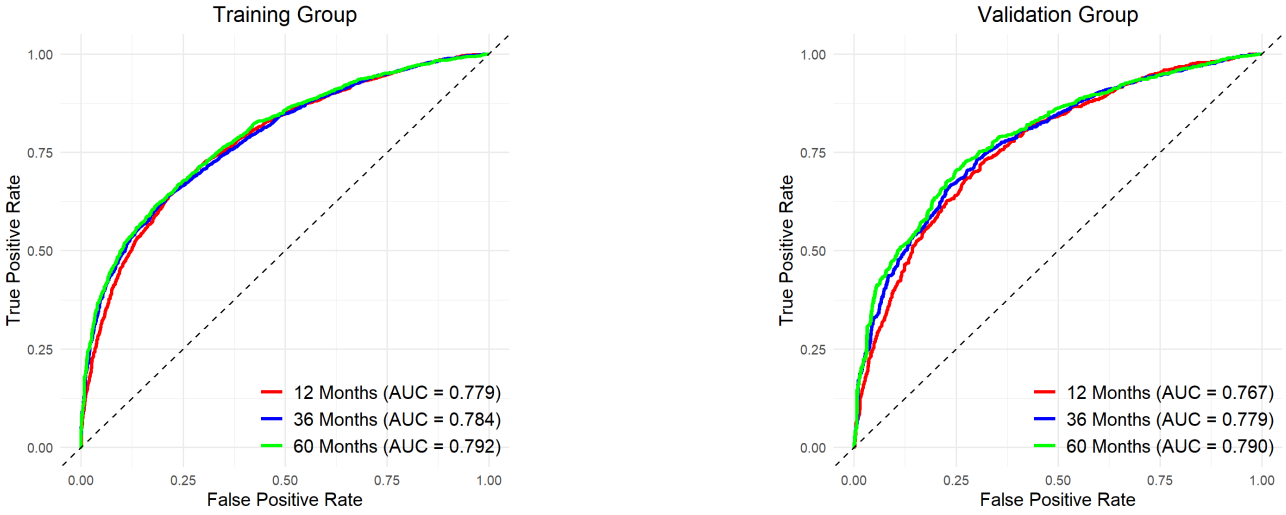
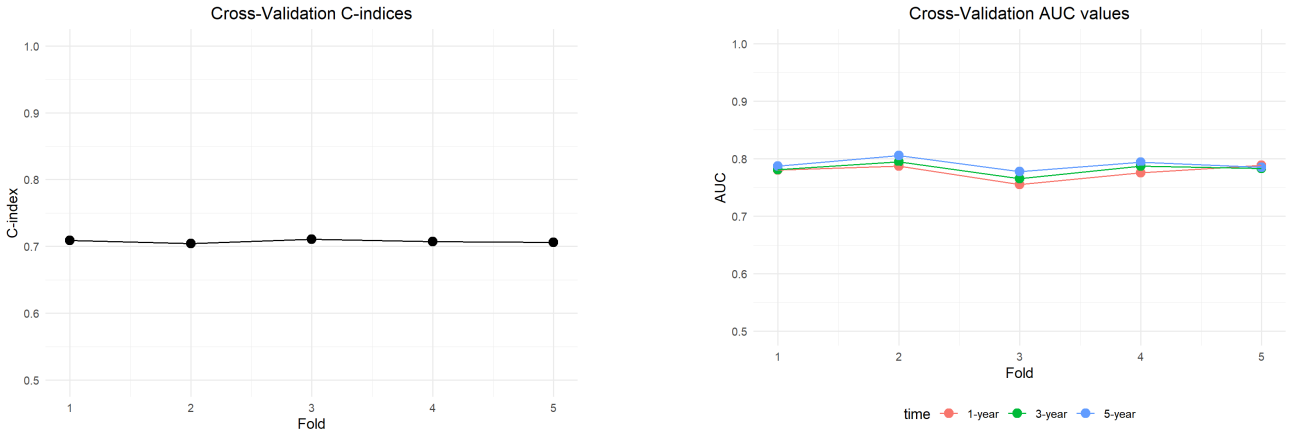
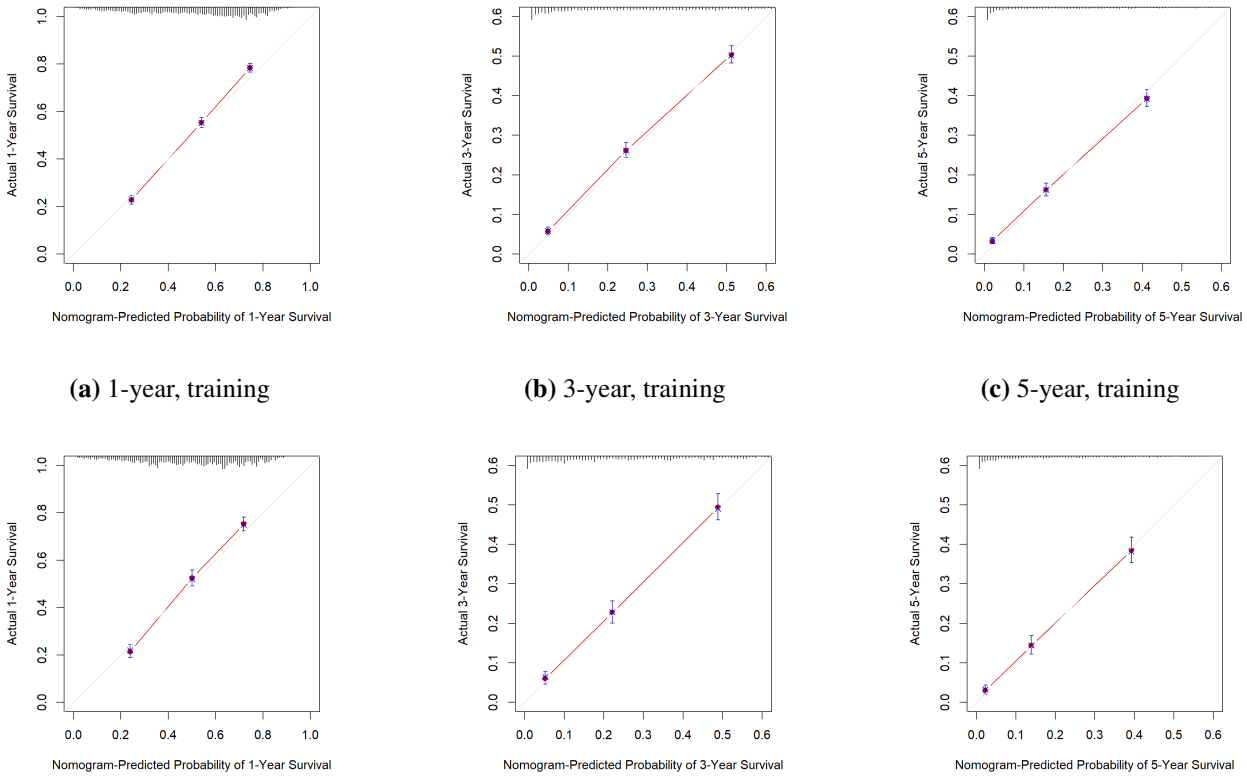


Figure 5. ROC curves of the nomogram in the training and validation groups.



(a) C-index **(b)** AUC values

Figure 6. Visualization of the C-index and AUC values from 5-fold cross-validation ($k = 5$).



(a) 1-year, training **(b)** 3-year, training **(c)** 5-year, training

(d) 1-year, validation **(e)** 3-year, validation **(f)** 5-year, validation

Figure 7. Calibration curves for the 1-, 3-, and 5-year predictions of the nomogram model.

A calibration curve is a graphical tool used to evaluate the accuracy of predictive models by comparing predicted probabilities of an event, such as disease risk, against the actual observed frequencies. The closer the curve aligns with the 45-degree ideal line, the better the model's

calibration, reflecting greater reliability in its predictions. This tool plays a critical role in assessing and refining a model's ability to produce well-calibrated risk estimates, thereby supporting informed clinical decision-making by minimizing over-optimistic or pessimistic predictions.

In both the training and validation groups, the calibration curves showed close agreement with the ideal diagonal, indicating strong consistency between predicted probabilities and observed outcomes. These findings further confirm the model's calibration accuracy. The corresponding plots are presented in Figure 7.

4. Development and validation of the RSF model

The RSF model is an ensemble learning method based on the bagging algorithm. As a classifier consisting of multiple decision trees, it enhances generalization performance by increasing diversity among individual learners. We fitted the RSF model to the training group and generated a VIMP measure, as illustrated in Figure 8.

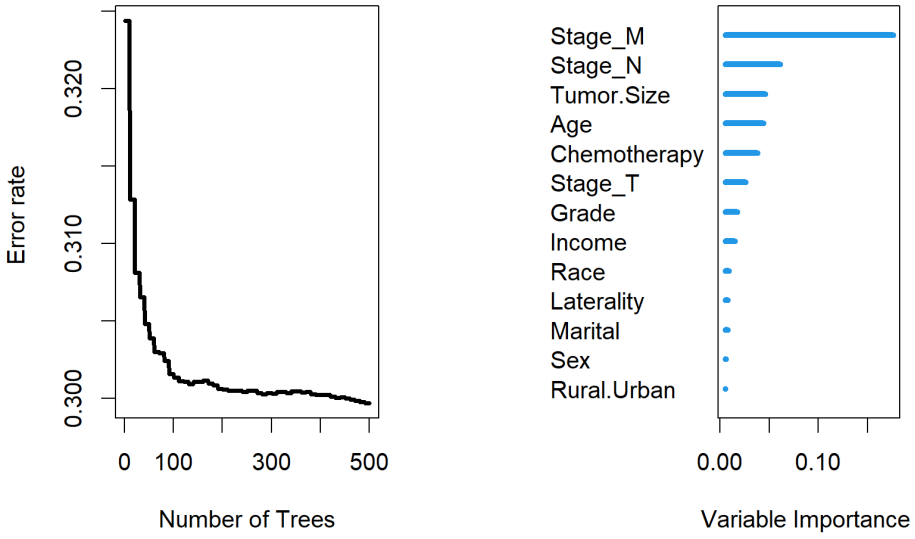


Figure 8. Performance and VIMP analysis of the RSF model.

As shown in Figure 8, the prediction error rate stabilized after 100 trees were built. We used a grid search to optimize the core hyperparameters of the RSF model, identifying the best combination as mtry = 2 (number of variables randomly sampled as candidates for splitting a node), and nodesize = 20 (minimum size of terminal nodes), which yielded optimal performance in the validation group. Using these parameters with 500 trees (ntree = 500), the model achieved a C-index of 0.733 in the training group and 0.698 in the validation group. VIMP analysis revealed that most covariates—except for region, gender, marital status, tumor laterality, and race—were significantly associated with survival and can be considered important prognostic factors.

Based on these prognostic variables, an RSF model was constructed. In line with the approach used

for the Cox model, patients were stratified into high- and low-risk groups using the median risk score. Kaplan-Meier survival analysis, coupled with log-rank testing, demonstrated statistically significant differences in survival between the risk groups in both the training and validation cohorts ($p < 0.0001$), with the high-risk group exhibiting markedly poorer prognosis. The model was also used to predict 1-, 3-, and 5-year overall survival in patients with lung cancer. ROC curve analysis confirmed favorable sensitivity and specificity within the training group.

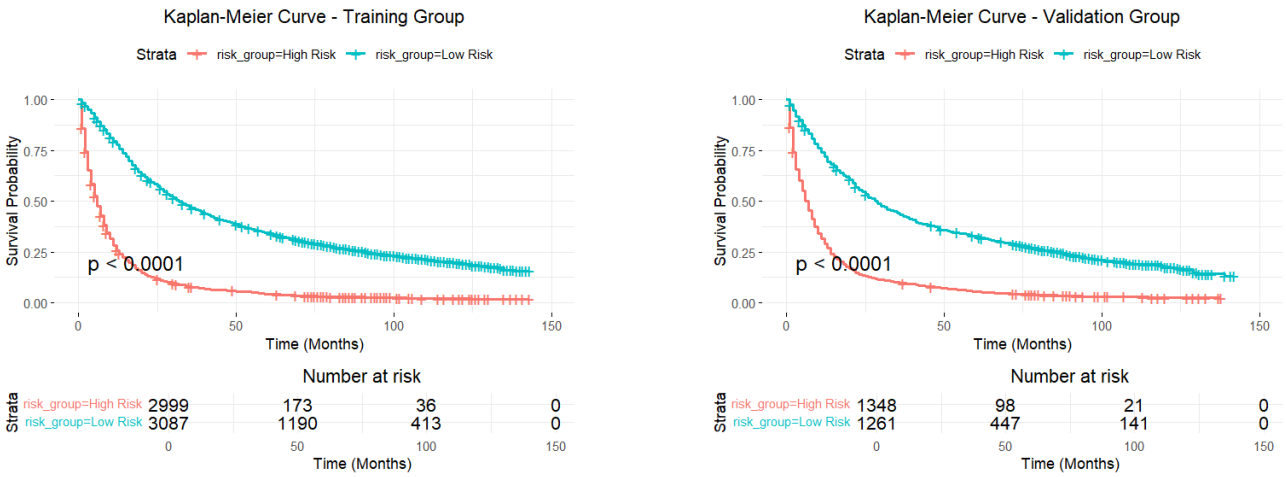


Figure 9. Kaplan-Meier survival curves for the two mortality risk subgroups in the training and validation groups.

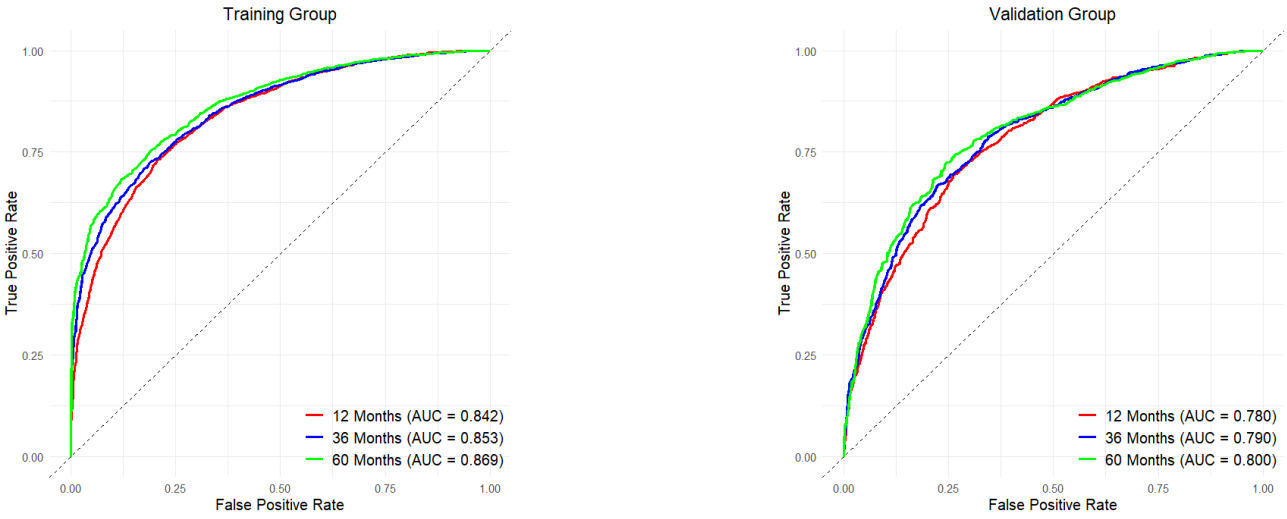


Figure 10. ROC curves of the RSF model in the training and validation groups.

In summary, the RSF model developed in this study effectively identified key prognostic factors and demonstrated robust performance in risk stratification for lung cancer patients. The model's ability to distinguish high-risk and low-risk subgroups with significant prognostic separation was consistently validated across both training and validation groups. Furthermore, its accurate prediction of 1-, 3-, and 5-year overall survival, as supported by ROC analysis, underscores the clinical utility of this RSF-

based approach. These results suggest that the RSF model serves as a reliable and non-parametric tool for prognostic prediction and could potentially inform personalized treatment strategies for patients with lung malignancies.

5. Conclusions

Lung cancer, also referred to as primary bronchogenic carcinoma, is a common malignant tumor with high clinical incidence [13, 14]. Despite recent progress in early screening and therapeutic strategies, it remains associated with high mortality and poor overall prognosis, underscoring the persistent need for reliable prognostic tools to facilitate personalized treatment planning. Existing survival prediction models predominantly rely on clinicopathological variables and conventional statistical approaches, with the Cox proportional hazards model long serving as the gold standard for cancer prognosis evaluation. However, advances in machine learning have led to the emergence of non-parametric models such as RSF, which have become increasingly prominent in survival analysis due to their superior flexibility and accuracy when handling complex, high-dimensional data [15, 16].

To compare these approaches, this study utilized clinical data from patients with lung cancer obtained from the SEER database between 2010 and 2018. Both Cox proportional hazards model and RSF model were employed to predict overall survival. Through this analysis, key independent prognostic factors, including age, tumor stage, and chemotherapy status, were identified, and the predictive performance of the two modeling approaches was systematically compared.

Regarding the Cox proportional hazards model, our findings are consistent with previous studies that identified similar prognostic factors in lung cancer. For example, in developing a prognostic model for survival in patients with non-small cell lung cancer, Liu et al. utilized Cox regression to incorporate variables including TNM stage, tumor grade, and the expression levels of specific genes [17]. Similarly, Hu et al. examined factors associated with extrathoracic metastasis at initial diagnosis in patients with small isolated lung cancers (tumor diameter ≤ 2 cm) using Cox regression, evaluating covariates, such as age, sex, histologic type, and primary tumor site, and subsequently developed a predictive nomogram [18].

The applicability of the Cox model relies on the proportional hazards assumption, which presumes linear covariate effects. In contrast, the RSF model demonstrates notable advantages in handling complex clinical datasets through its tree-based ensemble approach, effectively capturing nonlinear relationships and higher-order interactions among prognostic factors [19].

Consistent with these methodological strengths, our analysis shows that although the C-index of the RSF model is slightly lower than that of the Cox model, its time-dependent AUC values are consistently higher. This indicates superior predictive accuracy at specific time points and confirms its clinical utility. Our findings align with previous research. For example, a retrospective study by Xia et al., which compared four models (e.g., Cox proportional hazards and RSF) for predicting survival in lung adenocarcinoma patients, similarly reported that while the RSF model had a slightly lower C-index, it demonstrated a stronger capability for long-term prediction, particularly in scenarios involving complex variable interactions and nonlinear effects [20]

Beyond predictive performance, the two models also differed in their handling of prognostic variables. While the Cox model primarily incorporated clinicopathological factors, the RSF approach also identified meaningful contributions from socioeconomic variables such as income. Our findings

suggest that these factors may play important roles in lung cancer prognosis and warrant further investigation. This observation aligns with a meta-analysis by Finke et al., which reported a modest positive correlation between income and lung cancer survival, underscoring the potential value of integrating socioeconomic factors into personalized prognostic models, particularly in resource-limited settings [21].

Several limitations of this study should be acknowledged. The retrospective nature of the SEER data introduces potential selection bias, which may affect the external validity and generalizability of our findings. Although internal validation was conducted, external validation using independent prospective cohorts is still needed. Moreover, the database lacks detailed information on important prognostic influences such as targeted therapies, immunotherapies, smoking history, and environmental exposures, which may have considerably impacted survival outcomes [22, 23]. Future studies should therefore focus on large-scale, prospective designs to validate and refine these models.

In summary, this comparative analysis demonstrates that while the Cox model remains a foundational method in survival analysis, the RSF model offers superior predictive accuracy, especially in capturing complex, non-linear interactions among variables. The integration of machine learning techniques into clinical prognostic modeling shows considerable promise for enhancing individualized risk assessment and treatment planning. Future efforts should focus on expanding data sources to include more diverse clinical, socioeconomic, and lifestyle variables, alongside rigorous external validation, to improve the robustness and applicability of prognostic models in lung cancer [24].

Data availability

The data analyzed during the current study are available from the corresponding author on reasonable request.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The research was supported in part by the Natural Science Foundation of Henan Province (Grant No. 252300421487) and the Key Scientific Research Project of Henan Province for Colleges and Universities (Grant No. 24A910001).

Conflict of interest

The authors declare no potential conflict of interest with respect to the research, authorship and/or publication of this article.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 71: 209–249. https://doi.org/10.3322/caac.21660
2. Zhu J, Shi H, Ran H, Lai Q, Shao Y, Wu Q, (2022) Development and validation of a nomogram for predicting overall survival in patients with second primary small cell lung cancer after non-small cell lung cancer: A SEER-based study. *Int J Gen Med* 15: 3613–3624. https://doi.org/10.2147/IJGM.S353045
3. Alexander M, Wolfe R, Ball D, Conron M, Stirling RG, Solomon B, et al. (2017) Lung cancer prognostic index: A risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. *Br J Cancer* 117: 744–751. https://doi.org/10.1038/bjc.2017.232
4. Devaux A, Helmer C, Genuer R, Proust-Lima C, (2023) Random survival forests with multivariate longitudinal endogenous covariates. *Stat Methods Med Res* 32: 2331–2346. https://doi.org/10.1177/09622802231206477
5. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP, (2016) Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 44: 368–374. https://doi.org/10.1097/CCM.0000000000001571
6. Feng Y, Qiao H, Han X, Tang H, (2024) A prognostic nomogram of non-small cell lung cancer based on tumor marker inflammatory nutrition score. *Transl Lung Cancer Res* 13: 3392–3406. https://doi.org/10.21037/tlcr-24-708
7. Guo H, Nie G, Zhao X, Liu J, Yu K, Li Y, (2024) A nomogram for cancer-specific survival of lung adenocarcinoma patients: A SEER based analysis. *Surg Open Sci* 22: 13–23. https://doi.org/10.1016/j.sopen.2024.10.003
8. Tong Y, Cui Y, Jiang L, Pi Y, Gong Y, Zhao D, (2022) Clinical characteristics, prognostic factor and a novel dynamic prediction model for overall survival of elderly patients with chondrosarcoma: A population-based study. *Front Public Health* 10: 901680. https://doi.org/10.3389/fpubh.2022.901680
9. Kaplan EL, Meier P, (1985) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53: 457–481. https://doi.org/10.1080/01621459.1958.10501452
10. Efron B, Hastie T, (2016) *Computer Age Statistical Inference*, Cambridge: Cambridge University Press, 131–139. https://doi.org/10.1017/CBO9781316576533
11. Wolbers M, Koller MT, Witteman JC, Steyerberg EW, (2009) Prognostic models with competing risks: Methods and application to coronary risk prediction. *Epidemiology* 20: 555–561. https://doi.org/10.1097/EDE.0b013e3181a39056
12. Liu W, Zhou L, Zhao D, Wu X, Yue F, Yang H, et al. (2022) Development and validation of a prognostic nomogram in lung cancer with obstructive sleep apnea syndrome. *Front Med* 9: 810907. https://doi.org/10.3389/fmed.2022.810907

13. Bade BC, Dela Cruz CS, (2020) Lung cancer 2020: Epidemiology, etiology, and prevention. *Clin Chest Med* 41: 1–24. https://doi.org/10.1016/j.ccm.2019.10.001

14. Liao CY, Chen YM, Wu YT, Chao HS, Chiu HY, Wang TW, et al. (2024) Personalized prediction of immunotherapy response in lung cancer patients using advanced radiomics and deep learning. *Cancer Imaging* 24: 129. https://doi.org/10.1186/s40644-024-00779-4

15. Wang H, Shen L, Geng J, Wu Y, Xiao H, Zhang F, et al. (2018) Prognostic value of cancer antigen-125 for lung adenocarcinoma patients with brain metastasis: a random survival forest prognostic model. *Sci Rep* 8: 5670. https://doi.org/10.1038/s41598-018-23946-7

16. Roshanaei G, Safari M, Faradmal J, Abbasi M, Khazaei S, Factors affecting the survival of patients with colorectal cancer using random survival forest. *J Gastrointest Cancer* 53: 64–71. https://doi.org/10.1007/s12029-020-00544-3

17. Liu L, Shi M, Wang Z, Lu H, Li C, Tao Y, et al. (2018) A molecular and staging model predicts survival in patients with resected non-small cell lung cancer. *BMC Cancer* 18: 966. https://doi.org/10.1186/s12885-018-4881-9

18. Hu A, Chen Z, Liu C, Gao Y, Deng C, Liu X, (2022) Incidence and prognosis nomogram of small solitary lung cancer (≤2 cm) with extra-thoracic metastasis at initial diagnosis: A population-based study. *Cancer Control* 29: 10732748221141560. https://doi.org/10.1177/10732748221141560

19. Roshanaei G, Safari M, Faradmal J, Abbasi M, Khazaei S, (2022) Factors affecting the survival of patients with colorectal cancer using random survival forest. *J Gastrointest Cancer* 53: 64–71. https://doi.org/10.1007/s12029-020-00544-3

20. Xia K, Chen D, Jin S, Yi X, Luo L, (2023) Prediction of lung papillary adenocarcinoma-specific survival using ensemble machine learning models. *Sci Rep* 13: 14827. https://doi.org/10.1038/s41598-023-40779-1

21. Finke I, Behrens G, Weisser L, Brenner H, Jansen L, (2018) Socioeconomic differences and lung cancer survival-systematic review and meta-analysis. *Front Oncol* 8: 536. https://doi.org/10.3389/fonc.2018.00536

22. Yang S, Zhang Z, Wang Q, (2019) Emerging therapies for small cell lung cancer. *J Hematol Oncol* 12: 47. https://doi.org/10.1186/s13045-019-0736-3

23. Armstrong SA, Liu SV, (2018) Immune checkpoint inhibitors in small cell lung cancer: A partially realized potential. *Adv Ther* 36: 1826–1832. https://doi.org/10.1007/s12325-019-01008-2

24. Luo J, Hu J, Mulati Y, Wu Z, Lai C, Kong D, et al. (2024) Developing and validating a nomogram for penile cancer survival: A comprehensive study based on SEER and Chinese data. *Cancer Med* 13: e7111. https://doi.org/10.1002/cam4.7111