

Original Paper

A Machine Learning Approach to Predicting Mortality Risk in Chemotherapy-Treated Lung Cancer: Machine Learning Model Development and Validation

Jianjun Zou^{1,2,3*}, MD; Jinyi Huang^{4*}, MD; Katie Lu^{5*}, BD; Ao Lin^{4,6,7}, MD, PhD; Chen Xie⁴, MD; Jinrong Zhang⁴, MD; Boqi Rao^{4,6,8}, MD, PhD; Zhi Li⁴, MD; Dongming Xie⁴, MD; Ling Lu⁴, MD; Feng Luo⁴, BD; Jinbin Chen⁹, MD, PhD; Lei Yang^{4,10,11}, MD, PhD; Fuman Qiu⁴, MD, PhD; Xin Zhang^{6,7,12}, MD, PhD; Yibin Deng¹³, MD, PhD; Jiachun Lu^{4,10,11}, MD, PhD

¹Institute of Tuberculosis, Guangzhou Medical University, Guangzhou, Guangdong, China

²Guangzhou Key Laboratory of Tuberculosis Research, Guangzhou Chest Hospital, Guangzhou, Guangdong, China

³Department of General Medicine, Guangzhou Chest Hospital, Guangzhou, Guangdong, China

⁴School of Public Health, Guangzhou Medical University, Guangzhou, Guangdong, China

⁵School of Medicine, University of Arizona, Tucson, AZ, United States

⁶State Key Laboratory of Respiratory Disease, Guangzhou Medical University, Guangzhou, Guangdong, China

⁷Department of Thoracic Surgery, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China

⁸Emergency Department, The Second Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

⁹KingMed School of Laboratory Medicine, Guangzhou Medical University, Guangzhou, China

¹⁰The Key Laboratory of Advanced Interdisciplinary Studies, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China

¹¹The Institute for Chemical Carcinogenesis, Guangzhou Medical University, Guangzhou, Guangdong, China

¹²Guangzhou Institute of Respiratory Health, Guangzhou Medical University, Guangzhou, China

¹³Centre for Medical Laboratory Science, The Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, China

*these authors contributed equally

Corresponding Author:

Jiachun Lu, MD, PhD
School of Public Health
Guangzhou Medical University
Xinzao, Panyu District
Guangzhou, 511436
China
Phone: 86 20 37103043
Email: jclu@gzhmu.edu.cn

Abstract

Background: Accurately predicting the survival outcomes of patients with lung cancer receiving chemotherapy remains challenging.

Objective: To improve clinical management of this population, this study developed a multivariate machine learning (ML) model to assess all-cause mortality risk in chemotherapy-treated patients with lung cancer.

Methods: This study retrospectively recruited 1278 postchemotherapy patients with lung cancer from Guangzhou Chest Hospital between 2017 and 2019. Candidate features such as demographic characteristics, environmental exposures, clinical information, and patient-reported symptoms were collected via questionnaires and the electronic medical record system. The survival status and the deceased date were investigated twice a year. A total of 84 predictive models were constructed on the training set using 5 ML algorithms either individually or in pairwise combinations. The concordance index was used to identify the optimal model on the testing set, with performance validated via receiver operating characteristic curves, calibration curves, and decision curve analysis. Additionally, Shapley Additive Explanations and restricted cubic splines were applied for feature attribution analysis.

Results: The optimal model ultimately retained 21 prognosis-association features, including age, sex, BMI, smoking status, environmental smoke, the MD Anderson Symptom Inventory for Lung Cancer total score trajectories, cluster of differentiation 56, TNM stage, histology, and prechemotherapy blood biomarkers. On the testing set, the model acquired a concordance index of 0.702 (95% CI 0.652-0.753). The decision curves demonstrated positive clinical benefit when the risk thresholds were 0.40-0.69, 0.62-0.99, and 0.72-0.99 for 1-, 3-, and 5-year mortality predictions, respectively. The calibration curves showed that the predicted mortality probabilities fluctuated around the observed probabilities, and the Brier scores for 1-, 3-, and 5-year predictions were 0.20, 0.18, and 0.11, respectively. The area under the curve of the model was 0.740, 0.777, and 0.915 for 1-, 3-, and 5-year mortality predictions, respectively. Interpretability feature attribution analysis revealed that the significant features could predict all-cause mortality risk in chemotherapy-treated patients with lung cancer.

Conclusions: Our ML models exhibited acceptable discrimination, calibration, and clinical benefit in predicting the mortality risk of chemotherapy-treated patients with lung cancer, which could help clinicians in personalized prognostic management.

(*JMIR Med Inform* 2025;13:e72424) doi: [10.2196/72424](https://doi.org/10.2196/72424)

KEYWORDS

lung cancer; chemotherapy; mortality; machine learning; prediction model

Introduction

Lung cancer is the leading cause of cancer-related deaths globally. Lung cancer claimed approximately 1.8 million lives in 2022 alone, accounting for 18.7% of global cancer deaths [1]. With the advancements in medicine, the emergence of targeted and immune drugs introduces new options for lung cancer treatment. Personalized treatment strategies combining surgery, radiotherapy, chemotherapy, targeted therapy, and immunotherapy demonstrate significant efficacy in improving the living quality and long-term survival rates of patients with lung cancer. However, for some advanced lung cancers lacking driver mutations and with negative immune checkpoint expression, chemotherapy remains the primary treatment modality [2,3]. Thus, accurately predicting the mortality risk of chemotherapy-treated lung cancers to optimize treatment planning and patient care is of significant public health importance.

Machine learning (ML) models have become increasingly important tools for predicting the mortality risk of patients with lung cancer treated with chemotherapy, offering new insights and improved accuracy over traditional methods. While demographic and clinical characteristics are important, recent research has highlighted the potential benefits of incorporating additional data sources to enhance model performance [4-6]. For example, although the TNM stage is a recognized factor associated with the prognosis, it explains just less than 30% of the prognostic differences [7,8]. It is noted that genomic biomarkers present a means of refining prognosis assessment models, but the expensive genetic testing restricts the clinical application [9-11]. The above situation underscores the

requirement for a more effective and interpretable mortality risk stratification model for postchemotherapy patients with lung cancer.

The biopsychosocial framework advocates dissecting disease from diverse angles, including symptom burden. The MD Anderson Symptom Inventory for Lung Cancer (MDASI-LC), a patient-reported outcome (PRO) tool, offers key perspectives into the self-reported symptom experience of patients receiving chemotherapy [12]. Integrating longitudinal PRO information into prognostic models has notable value in refining the prediction of patient outcomes [13-16].

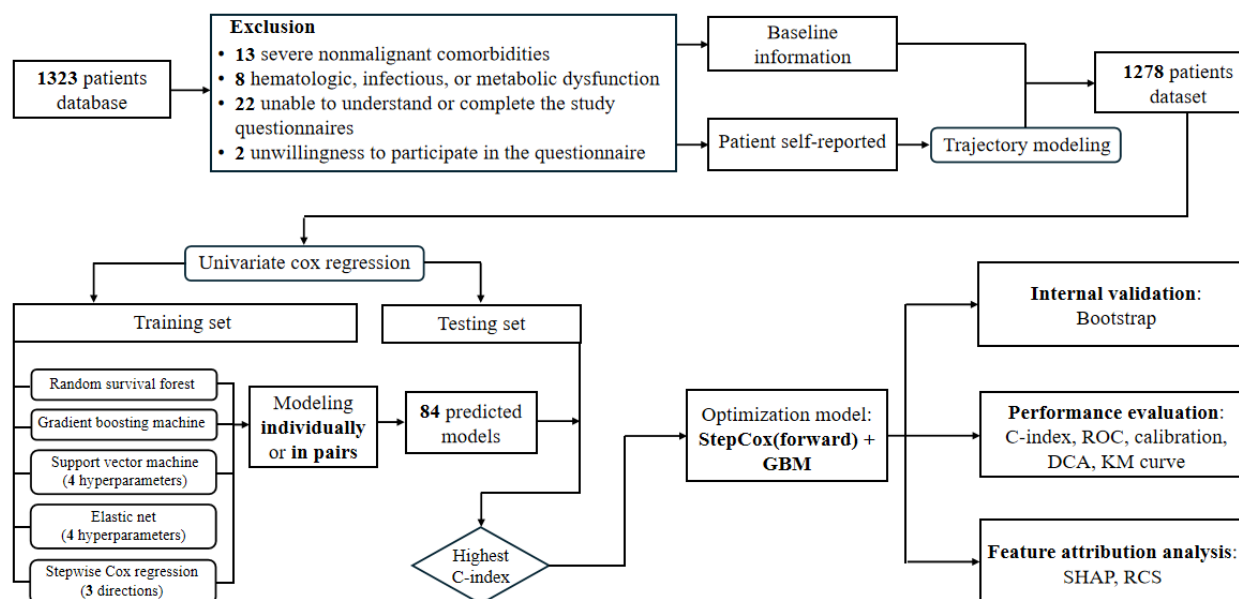
This study adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis guidelines as much as possible to ensure the transparent and accurate reporting of our predictive model development and validation [17]. This study developed a multivariable ML model that integrated demographic characteristics, clinical features, environmental factor exposure, immunohistochemical indicators, and total PRO score to predict the mortality risk of chemotherapy-treated patients with lung cancer. The model aims to provide clinicians with an effective tool for predicting the risk of death in patients with lung cancer after chemotherapy, assisting clinicians in customizing treatment plans, optimizing patient care plans, and improving patient prognosis.

Methods

Overview

An overview of the entire study design and ML framework is provided in [Figure 1](#).

Figure 1. The flowchart of this study. This study provided a comprehensive framework for building machine learning models to predict mortality risk for patients with lung cancer after chemotherapy. C-index: concordance index; DCA: decision curve analysis; GBM: gradient boosting machine; KM: Kaplan-Meier; RCS: restricted cubic spline; ROC: receiver operating characteristic curve; SHAP: Shapley Additive Explanations.



Study Population

This retrospective and semilongitudinal investigation included 1278 patients with lung cancer who received platinum-based chemotherapy at Guangzhou Chest Hospital between November 2017 and 2019. Demographic characteristics, environmental exposures, clinical information, and patient-reported symptoms were collected using questionnaires and the electronic medical record system. Patients' TNM stage was identified by the seventh edition of the classification system.

Patients who met the following inclusion criteria were included (1) histologically confirmed diagnosis of primary lung cancer, (2) absence of prior chemotherapy for lung carcinoma, and (3) demonstrated ability to comprehend and complete study questionnaires with clear consciousness.

Patients with the following presentations that would bias their symptom reporting or outcome were excluded: (1) severe nonmalignant comorbidities; (2) hematopoietic, infectious, or metabolic dysfunction; and (3) preexisting psychological disorders. The specific number of excluded participants is presented in the flowchart of Figure 1.

Patient-Reported Symptom

Patients completed self-report questionnaires at admission and the first 4 chemotherapy cycles. The professional questionnaires included (1) the MDASI-LC: this is a validated PRO tool specifically designed to assess symptom severity in patients with lung cancer. It consists of 22 items divided into 3 parts: core symptoms, lung cancer-specific symptoms, and additional symptoms. Each item is scored on a scale from 0 to 10, with higher scores indicating more severe symptoms [18]. (2) The Karnofsky Performance Status (KPS): this is a widely used tool for assessing the overall health status and functional state of patients. The KPS score ranges from 0 to 100, with higher scores indicating better functional status. It is commonly used in

clinical practice to evaluate patient prognosis and treatment response, and it has been validated in various studies [19]. (3) The Zubrod Performance Score (ZPS): this tool assesses the physical condition of patients, with scores ranging from 0 to 5. Higher scores indicate worse physical status. ZPS is another widely used tool in clinical practice, providing valuable insights into patients' daily living abilities and overall health [20].

The Primary End Point

All-cause mortality served as the primary end point in this study. The professionals monitored the patients biannually by telephone. The death date was recorded if the patients had died; otherwise, they would be followed until death. The time horizon for this outcome is defined as 1-, 3-, and 5-year posttreatment initiation. The follow-up period in this research ranged from 1 month to 11 years.

Statistical Analysis Methods

The normality of continuous variables was assessed using the Shapiro-Wilk test, and the group comparisons were made using the 2-tailed *t* test or the Mann-Whitney *U* test. Categorical variables were analyzed using the chi-square or the Fisher exact test. Survival curves were generated using Kaplan-Meier and compared using the log-rank test. Hazard ratio and 95% CI were estimated using Cox regression. Statistical significance was defined when $P < .05$ (2-tailed). Missing data imputation was performed using the missForest algorithm, which accommodated mixed data types and handled nonlinearities and outliers.

The reliability of the scales was assessed using Cronbach α to measure internal consistency. Pearson correlation coefficients were calculated to examine the correlations between the 3 scales, assessing their validity. Longitudinal symptom burden trajectories were identified using group-based trajectory modeling (GBTM), with data from the 3 scales assessed across 5 separate time points [21]. GBTM criteria means odds of correct

classification >5 , average posterior probability (Avepp) >0.7 , posterior probability of group membership (P_j) $>5\%$, correspondence between P_j and probability of group membership (π_j), relative entropy (E_k) >0.7 , and a minimized Bayesian information criterion value.

ML Methods

Random survival forest (RSF), gradient boosting machine (GBM), survival support vector machine (survival-SVM), elastic net (Enet), and stepwise Cox regression were used to develop a prognostic model [22-25]. For the RSF and GBM models, the best hyperparameters for these 2 models were determined by constructing separate base models on the training set and performing preliminary cross-validation and performance evaluation. For the hyperparameter γ of the survival-SVM model, the hyperparameter α of the Enet model, and the 3 directions of the stepwise Cox regression model, a grid search method was used. The specific parameters of the five algorithms are as follows: (1) RSF: ntree=1000, nodesize=32, mtry=10; (2) GBM: ntrees=3750, interaction.depth=2, n.minobsinnode=20; (3) survival-SVM: γ in 0.25, 0.5, 1, 2, 4; (4) Enet: α in 0.1, 0.2, 0.3, 0.4; and (5) stepwise Cox regression: direction in both, backward, forward. Five separate ML models with a total of 13 models (different hyperparameter models of the same algorithm are considered as different models) were first constructed: RSF (1 model), GBM (1 model), survival-SVM (4 hyperparameters and 4 models), Enet (4 hyperparameters and 4 models), and stepwise Cox regression (3 directions and 3 models). Second, two-by-two combinations of the above 13 models in sequential order totaled 71 models, for a total of 84 models being constructed.

Continuous features were not binarized or split-boxed during the modeling process in order to maximize the amount of information retained in the data, and for categorical features, such as sex, smoking status, and tumor staging, the modeling process used discrete values.

Model discrimination, accuracy, and clinical benefit were evaluated using the concordance index (C-index), decision curve analysis, calibration curves, and the area under the curve (AUC). Internal validation of the models was performed using the bootstrap method. Feature attribution analysis was conducted using Shapley Additive Explanations and restricted cubic splines (RCSs).

Ethical Considerations

The protocol for this study was approved by the ethics review board of Guangzhou Medical University. All patients were informed of the data collection methods and privacy protection measures. All patient identifiers were removed after data collection and replaced with unique nonidentifiable serial numbers, and deidentified data were stored in a password-protected electronic database, with access restricted to the core research team. The deidentified data were only used for this study without sharing with third parties without additional approval.

Results

Baseline Characteristics of the Patients

This study enrolled 1278 eligible patients, with a median age of 61 (IQR 54-68) years, comprising 946 (74%) male and 332 (26%) female patients. Among 1278 enrolled patients with lung cancer undergoing chemotherapy, 408, 94, and 19 cases were alive at the 1-, 3-, and 5-year follow-ups, yielding overall survival rates of 31.9% (95% CI 29.4%-34.6%), 7.4% (95% CI 6%-9%), and 1.5% (95% CI 0.9%-2.4%), respectively. The specific distribution of histological types in our study population is as follows: adenocarcinoma (876/1274, 68.8%), squamous cell carcinoma (247/1274, 19.4%), small cell lung cancer (SCLC; 125/1274, 9.8%), and other types (26/1274, 2%). Some of the more important characteristics can be found in [Table 1](#), and the complete [Table 1](#) can be found in [Multimedia Appendix 1](#).

Table 1. Baseline characteristics of chemotherapy-treated patients with lung cancer.

Variable	Missing, n (%)	Baseline data (N=1278)		Statistics	P value
		Living (n=758)	Deceased (n=520)		
Sex, n (%)	0 (0)			$\chi^2_1=0.6$.43
Male		555 (73.2)	391 (75.2)		
Female		203 (26.8)	129 (24.8)		
Age (years), median (IQR)	0 (0)	60.0 (53.0-66.0)	63.0 (57.0-70.0)	$z=5.5$	<.01 ^a
BMI (kg/m ²), median (IQR)	48 (3.8)	20.8 (18.9-23.0)	20.6 (18.4-22.6)	$z=-2.1$.03 ^a
Education, n (%)	80 (6.3)			7.1 (2)	.03
Never attended school or primary school		377 (53.4)	294 (59.8)		
Junior high school or high school		304 (43.1)	175 (35.6)		
University and above		25 (3.5)	23 (4.7)		
WBC ^b (10 ⁹ per liter), median (IQR)	150 (11.7)	8 (6-10)	8 (7-11)	$z=2.9$	<.01 ^a
HGB ^c (g/L), median (IQR)	150 (11.7)	125.0 (111.0-136.0)	121.0 (107.0-134.0)	$z=-3.4$	<.01 ^a
D-dimer (mg/L), median (IQR)	150 (11.7)	0.69 (0.34-1.87)	0.91 (0.44-2.52)	$z=3.7$	<.01 ^a
CEA ^d (ng/mL), median (IQR)	150 (11.7)	5 (2-20)	7 (3-52)	$z=4.6$	<.01 ^a
CA12-5 ^e (U/mL), median (IQR)	150 (11.7)	31.2 (14.7-100.8)	52.9 (19.3-145.7)	$z=4.1$	<.01 ^a
CA19-9 ^f (U/mL), median (IQR)	150 (11.7)	13.0 (7.0-26.3)	16.3 (7.2-44.9)	$z=2.8$	<.01 ^a
CYFRA21-1 ^g (ng/mL), median (IQR)	150 (11.7)	5 (3-11)	8 (4-18)	$z=7.1$	<.01 ^a
Histology, n (%)	4 (0.3)			$\chi^2_3=1.0$.79
LUAD ^h		522 (69.1)	354 (68.2)		
LUSC ⁱ		149 (19.7)	98 (18.9)		
SCLC ^j		69 (9.1)	56 (10.8)		
Others		15 (2.1)	11 (2.1)		
Tumor stage, n (%)	31 (2.4)			$\chi^2_1=1.5$.22
T1+T2		281 (38)	176 (34.7)		
T3+T4		458 (62)	332 (65.4)		
Nodal stage, n (%)	68 (5.3)			$\chi^2_3=16.7$	<.01
N1		85 (11.9)	30 (6.1)		
N2		69 (9.6)	34 (6.9)		
N3		296 (41.3)	212 (42.9)		
N4		266 (37.2)	218 (44.1)		
Metastasis stage, n (%)	156 (12.2)			$\chi^2_1=28.3$	<.01
M1		174 (26.1)	59 (13)		
M2		493 (73.9)	396 (87)		

^aUsing the Mann-Whitney *U* test.^bWBC: white blood cell.^cHGB: hemoglobin.^dCEA: carcinoembryonic antigen.^eCA12-5: carbohydrate antigen 12-5.^fCA19-9: carbohydrate antigen 19-9.

^gCYFRA21-1: recombinant cytokeratin fragment antigen 21-1.
^hLUAD: lung adenocarcinoma.
ⁱLUSC: lung squamous cell carcinoma.
^jSCLC: small cell lung cancer.

Patient Report Outcomes Trajectory Modeling

GBTM was applied to characterize PRO trajectories. Table 2 presents detailed information on the GBTM evaluation indicators and trajectory groups. A comprehensive evaluation process is described in Tables S1-S6 in Multimedia Appendix 2. In particular, following a rigorous evaluation of model fit indices and clinical interpretability, 3 distinct trajectory groups were

identified for the MDASI-LC total score (group A: high burden, group B: medium burden, and group C: low burden), 3 for the KPS score (group A: high burden, group B: medium burden, and group C: low burden), and 2 for the ZPS score (group A: high burden and group B: low burden). All equation coefficients showed significant differences ($P<.001$; Figure S1 in Multimedia Appendix 2).

Table 2. The number of each scale trajectory groups, classification counts, and results of various evaluation indicators for the scale^a.

Scales	Group number	OCC ^b	Avepp ^c (%)	P_j^d (%)	π_j^e (%)	E_k^f	BIC ^g
MDASI-LC ^h	3 (3 0 3)	33.2/33.4/81.1	98.24/90.83/93.18	62.82/22.89/14.29	62.71/22.87/14.42	0.91	-9140.61
KPS ⁱ	3 (0 0 0)	943.70/43.10/60.80	98.03/95.25/99.05	4.76/32.60/62.67	5.01/31.74/63.25	0.95	-6565.35
ZPS ^j	2 (0 2)	5.60/538.10	93.74/98.94	90.84/9.16	85.26/14.74	0.77	-376.79

^aOptimization model criteria: The odds of correct classification should surpass a value of 5, reflecting a high likelihood of accurate classification. Each group should exhibit an Avepp exceeding the threshold of 0.7, indicating a high degree of confidence in the classification. The P_j should be greater than 5%, signifying that the probability of an entity belonging to a particular group is statistically significant. There should be a close alignment between P_j and the π_j , ensuring that the posterior probabilities are congruent with the a priori probabilities. The E_k should be greater than 0.7, indicating a substantial degree of differentiation between the groups in terms of information content.
^bOCC: odds of correct classification.
^cAvepp: average posterior probability.
^d P_j : posterior probability of group membership.
^e π_j : probability of group membership.
^f E_k : relative entropy.
^gBIC: Bayesian information criterion.
^hMDASI-LC: MD Anderson Symptom Inventory for Lung Cancer.
ⁱKPS: Karnofsky Performance Status.
^jZPS: Zubrod Performance Score.

Univariate Cox Regression Analysis of Feature Selection

The missForest algorithm did not induce statistically significant changes in the data distribution after baseline imputation. A detailed comparison of the baseline data distribution before and after imputation is reported in Table S7 in Multimedia Appendix 2.
Univariate Cox regression analysis (P value of $<.05$ is statistically significant) identified 27 features significantly associated with prognosis in patients with lung cancer receiving chemotherapy, which served as an initial feature selection step. A comprehensive summary of the univariate analysis is available in Table S8 in Multimedia Appendix 2. Beyond statistically significant findings, an additional 10 features with perceived prognostic relevance, such as a family history of respiratory disease, were included in feature selection despite their lack of

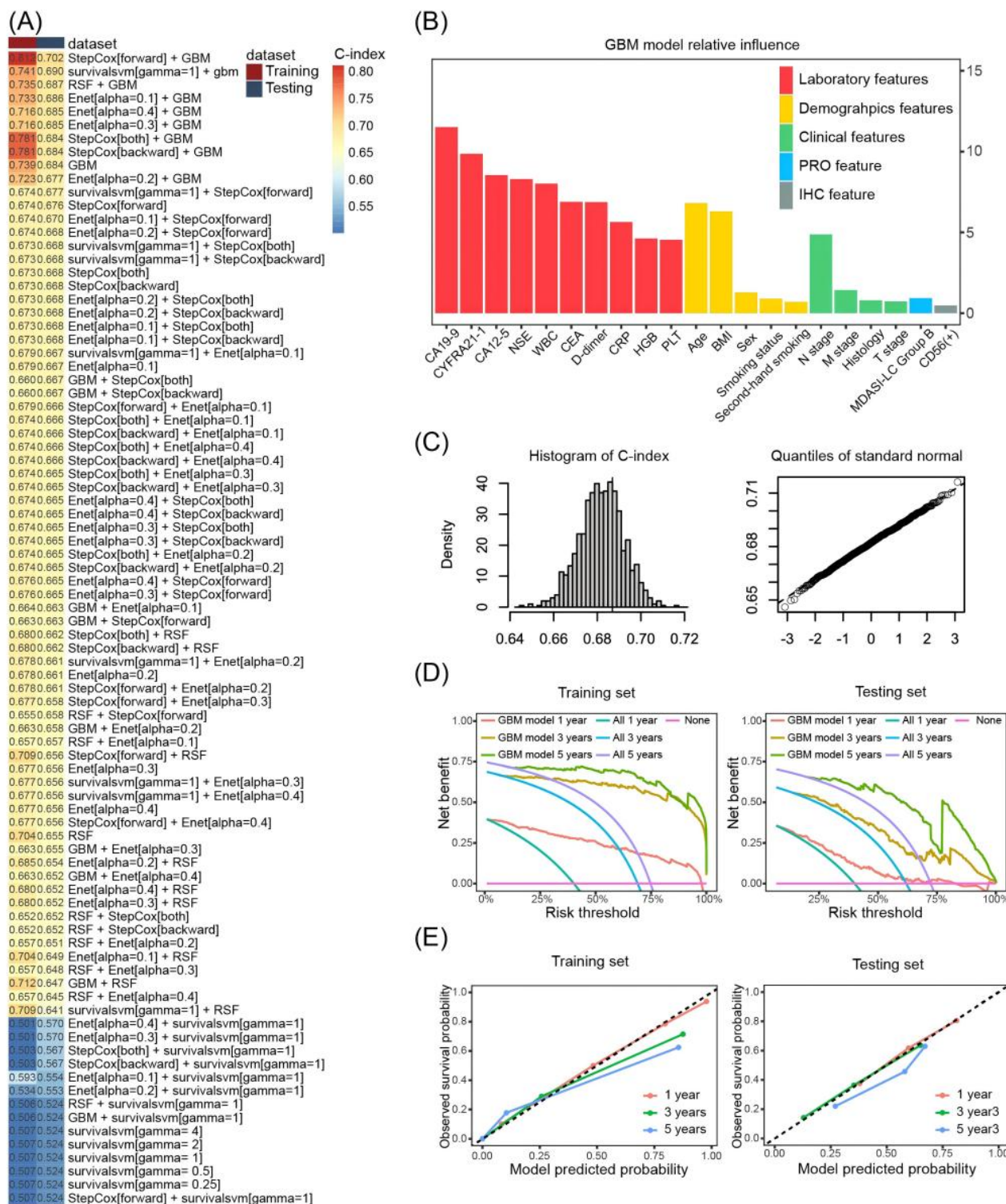
statistical significance in the univariate analysis. These features were retained to ensure a comprehensive evaluation of potential prognostic factors and to allow the ML models to capture complex interactions between features. Finally, 37 features were retained for the ML model development.

Models Development and Internal Validation

The dataset of 1278 observations was randomly split into a training dataset ($n=895$) and a testing dataset ($n=383$) at a 7:3 ratio. Table S9 in Multimedia Appendix 2 shows that there is no difference between the distributions of the main baseline features of the patients in the training and testing sets. On the training set, a total of 84 predicted models were built with 5 ML algorithms individually or binately 5 ML models were trained and then tested both individually and in all possible combinations of pairs, yielding a total of 84 predictive models (Figure 2A).



Figure 2. Illustration of the machine learning procedure for prediction model development and validation. (A) displays a ranking of 84 predictive models from highest to lowest based on their C-index calculated on the testing set. These models resulted from all possible 2-way combinations of 5 machine learning models. (B) presents the ranked feature importance identified by the GBM model, with the y-axis indicating the relative importance. (C) displays the C-index distribution along with results of a Shapiro-Wilk normality test, obtained during an internal validation using bootstrap resampling. (D) and (E) display DCA curves and calibration plots, respectively, for predicting 1-, 3-, and 5-year mortality risk using the optimal model on both the training and the testing set. CA12-5: carbohydrate antigen 12-5; CA19-9: carbohydrate antigen 19-9; CD56: cluster of differentiation 56; CEA: carcinoembryonic antigen; C-index: concordance index; CRP: C-reactive protein; CYFRA21-1: recombinant cytokeratin fragment antigen 21-1; DCA: decision curve analysis; Enet: elastic net; GBM: gradient boosting machine; HGB: hemoglobin; IHC: immunohistochemistry; MDASI-LC: MD Anderson Symptom Inventory for Lung Cancer; NSE: neuron-specific enolase; PLT: platelet count; PRO: patient-reported outcome; RSF: random survival forest; WBC: white blood cell.



On the testing set, we calculated the C-index of the 84 predicted models to identify the optimal predictive model. We found that

the optimal predictive model was a combination of forward stepwise Cox regression and GBM and demonstrated the highest

C-index of 0.702 (95% CI 0.652-0.753; [Figure 2A](#)). The final feature set of the optimal predictive model comprised 5 demographic features, 10 prechemotherapy laboratory features, 1 PRO feature, 4 clinical features, and 1 immunohistochemical feature ([Figure 2B](#)).

In total, 1000 bootstrap resamplings were performed and yielded an average C-index of 0.682 (95% CI 0.676-0.689) on the testing dataset. Shapiro-Wilk test indicated that the distribution of the C-index was normal ($P=.86$; [Figure 2C](#)). The evidence suggested a reasonable robustness of the optimal predictive model.

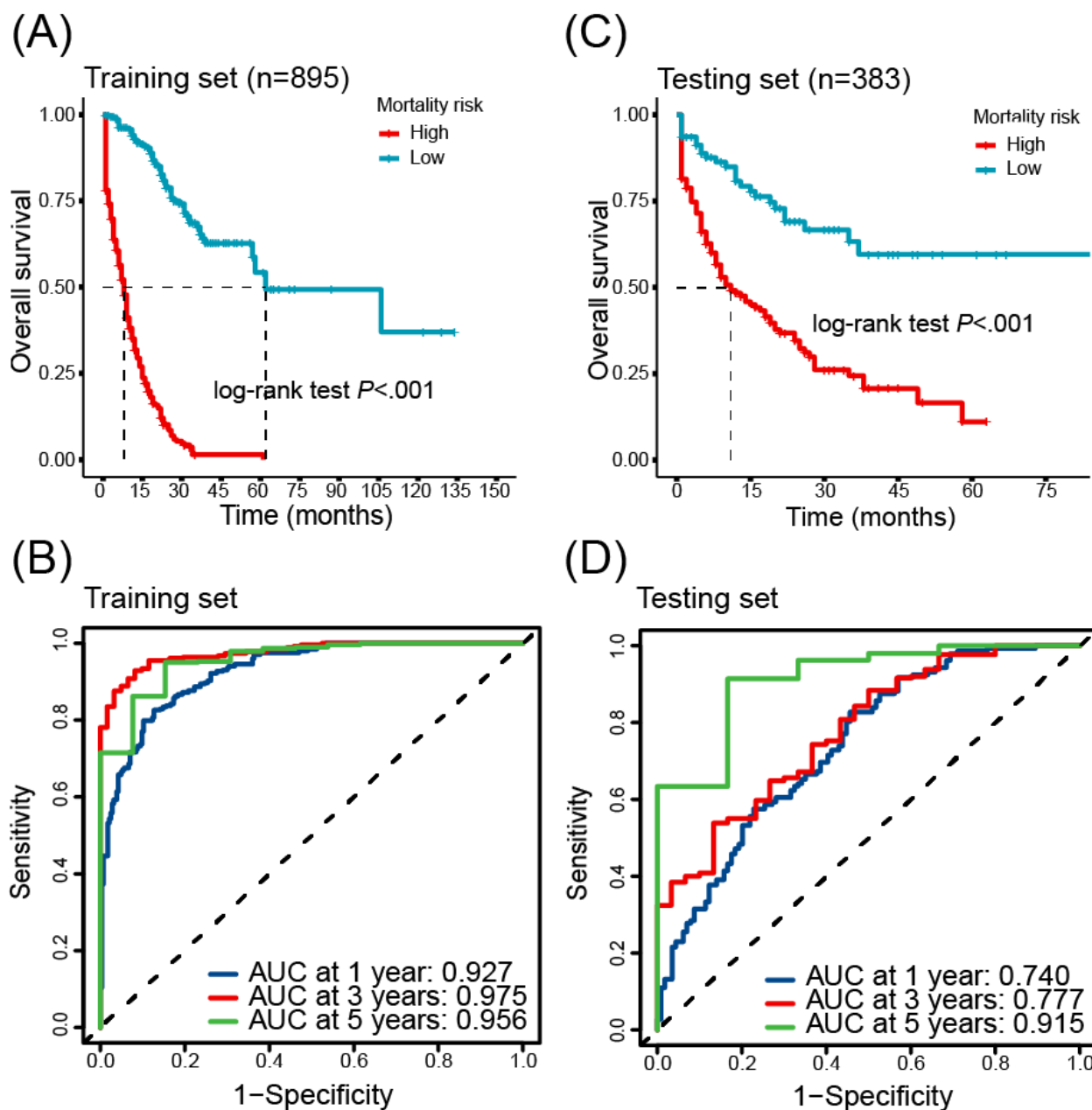
Performance Evaluation of the Model

Performance evaluation of the model was implemented on both the training and testing sets. The decision curve analysis demonstrated that the predicted model had a net benefit over the strategies of predicting all deaths or all survivals when the risk thresholds were 0.40-0.97, 0.68-0.99, and 0.75-0.99 for 1-, 3-, and 5-year mortality predictions, respectively, on the training set, while 0.40-0.69, 0.62-0.99, and 0.72-0.99 on the testing set ([Figure 2D](#)).

In addition, the calibration curves displayed strong agreement between model predictions and observed outcomes ([Figure 2E](#)). The Brier scores for 1-, 3-, and 5-year mortality predictions were 0.11, 0.07, and 0.08, respectively, on the training set, while 0.20, 0.18, and 0.11 on the testing set.

Furthermore, we determined an optimal risk score cutoff point by the *survminer* package to stratify patients into high-risk and low-risk groups. Kaplan-Meier survival analysis showed a significant difference in survival curves between the 2 groups, indicating that the model could well stratify patients when predicting 1-, 3-, or 5-year survival ($P<.01$; [Figures 3A and C](#)). The time-dependent receiver operating characteristic curve analysis found that the model demonstrated AUC values of 0.927, 0.975, and 0.956 for 1-, 3-, and 5-year mortality predictions, respectively, on the training set, while 0.740, 0.777, and 0.915 on the testing set, which exhibited robust predictive accuracy for both short-term and long-term mortality risk ([Figures 3B and D](#)).

Figure 3. Evaluation of the mortality risk prediction model's prognostic performance. (A) and (C) show the Kaplan-Meier curve for high- and low-risk patients, generated from the training and testing set, respectively, illustrating significant differences in survival. (B) and (D) present the time-dependent receiver operating characteristic curves assessing 1-, 3-, and 5-year mortality prediction in the training and testing data, respectively. AUC: area under the curve.

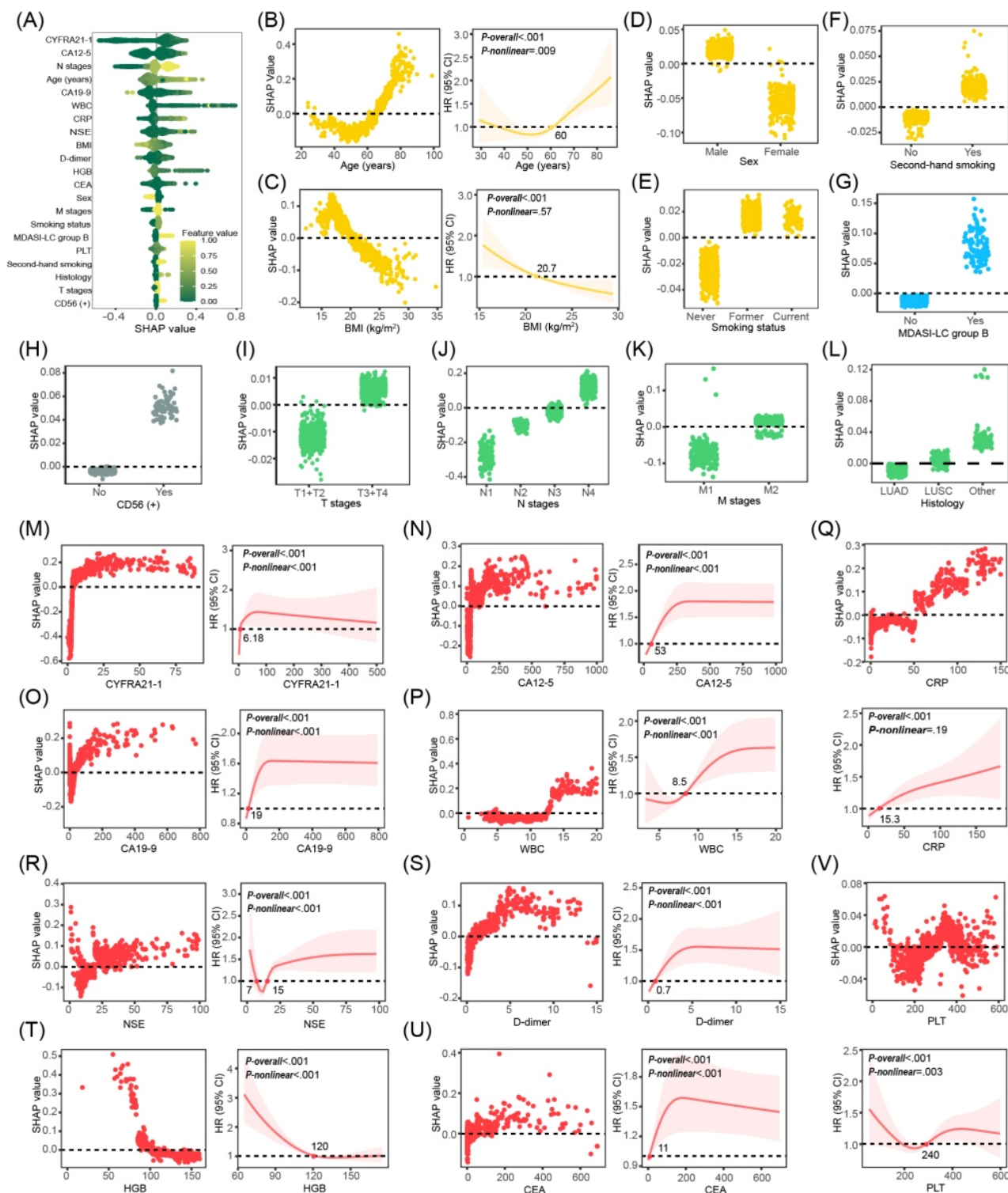


Feature Attribution Analysis

Figure 4A provides insights into the most significant features influencing the model's prediction of mortality risk in this study. These insights facilitate an improved understanding of the

biological basis of risk and the identification of potential targets to optimize clinical care. The visualization of these influential features highlights their importance in guiding clinical decisions and the development of personalized treatment approaches grounded in individual patient risk profiles.

Figure 4. Visualization of feature importance and variable relationships influencing mortality risk. (A) displays the global feature importance. (B-F) illustrate partial dependence plots for demographic characteristics and RCS analyses for continuous variables. (G and H) and (I-L) show partial dependence plots for symptom burden PRO trajectories, immunohistochemical indicators, and tumor biological indicators, respectively. (M-V) provide partial dependence plots and RCS analysis results for prechemotherapy blood biomarkers. All partial dependence plots are annotated with the zero-SHAP value position, while the RCS plots show the points where changes in mortality risk occur. CA12-5: carbohydrate antigen 12-5; CA19-9: carbohydrate antigen 19-9; CD56: cluster of differentiation 56; CEA: carcinoembryonic antigen; CRP: C-reactive protein; CYFRA21-1: recombinant cytokeratin fragment antigen 21-1; HGB: hemoglobin; HR: hazard ratio; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; MDASI-LC: MD Anderson Symptom Inventory for Lung Cancer; NSE: neuron-specific enolase; PLT: platelet count; RCS: restricted cubic spline; SHAP: Shapley Additive Explanations; WBC: white blood cell.



The use of RCS in this study facilitated the detection of key inflection points within continuous variables, thereby providing valuable insights into the intricate relationships among factors

that contribute to mortality risk in patients with lung cancer. This refined understanding can support clinical decision-making and provide personalized approaches to patient management.

Age emerged as a significant predictor of mortality risk, with a key threshold at 60 years. Individuals younger than 60 years of age presented with a lower and less prominent mortality risk, while the risk accelerated beyond this age (Figure 4B). BMI demonstrated an inverse relationship with mortality, with a critical value of 20.7 kg/m², below which mortality risk was diminished (Figure 4C). Furthermore, male sex, smoking history (inclusive of both former and current smoking), and passive smoke exposure were all linked to an increased mortality risk (Figures 4D and F).

Patients in the MDASI-LC relatively high-symptom burden group exhibited a higher mortality risk, with a median symptom score of 45.4 (IQR 36.2-51.4), while the other group had a score of 8.2 (IQR 5.2-19.8; Figure 4G). A higher chemotherapy symptom burden was associated with reduced survival.

Key clinical features, including the presence of cluster of differentiation 56 (CD56) positivity, T3 or T4 stage, N2, N3, and N4 stages, and M2 stage, were associated with a significantly higher relative risk of mortality. No significant difference in mortality risk was observed between patients with lung squamous cell carcinoma and lung adenocarcinoma, while patients with SCLC or other less frequently occurring lung cancers demonstrated decreased survival times (Figures 4H-L).

Prechemotherapy levels of laboratory markers recombinant cytochrome fragment antigen 21-1, carbohydrate antigen 12-5, carbohydrate antigen 19-9, white blood cell count, D-dimer, and carcinoembryonic antigen showed a nonlinear increase in mortality risk, with respective critical values identified at 6.18 ng/mL, 53 U/mL, 19 U/mL, $8.5 \times 10^9/L$, 0.7 mg/L, and 11 ng/mL, respectively. Mortality risk was significantly higher beyond these thresholds. Hemoglobin exhibited a distinct pattern, with mortality risk increasing significantly at levels below 120 g/L. In contrast, C-reactive protein was associated with a linear increase in mortality risk, with a risk inflection point at 15.3 mg/L. Prechemotherapy neuron-specific enolase levels exhibited a nonlinear relationship with mortality, where risk increased both below 7 ng/mL and above 15 ng/mL. Platelet counts also had a nonlinear relationship with mortality risk, displaying an initial decrease followed by an increase, with a turning point at 240×10^9 per liter (Figures 4M-V).

The observed relationships are visually depicted in Shapley Additive Explanations partial dependence plots and through RCS analyses (Figures 4B-V). Subgroup analyses further confirmed these observations, demonstrating statistically significant differences in mortality risk associated with the described characteristics ($P < .05$), except for prechemotherapy platelet levels (Figure S2 in Multimedia Appendix 2 illustrating subgroup analysis validation performed after determining key values via feature attribution analysis).

Discussion

Principal Findings

This study developed a combined ML prediction model by integrating multisource data to assess the all-cause mortality risk of patients with lung cancer receiving chemotherapy. It

identified key prognosis-related factors, including demographic characteristics, prechemotherapy laboratory indicators, longitudinal symptom burden, and clinical features, and analyzed the correlation patterns between each factor and mortality risk. The model exhibited favorable performance in discriminative ability, predictive accuracy, and clinical utility, with overall satisfactory effectiveness.

Demographic and Lifestyle Impacts on Lung Cancer Prognosis

Age is a core factor influencing the chemotherapy outcomes of patients with lung cancer. Aging is accompanied by decreased genomic and transcriptomic stability, leading to the accumulation of mutations and abnormal gene expression, which promotes the development of malignant tumor phenotypes and reduces treatment sensitivity. Meanwhile, aging-induced metabolic disorders (such as abnormal glucose metabolism and insulin resistance) provide a favorable environment for tumor growth; the decline in immune function weakens the body's antitumor capacity; and the reduction in tissue repair ability increases the risk of chemotherapy side effects, ultimately affecting survival [26-28]. The impact of sex on prognosis may be related to the regulation of tumor progression and chemotherapy response by sex hormones (especially estrogen), but the specific mechanism requires further research [29-31]. BMI showed a negative correlation with mortality risk, with a critical value of 20.7 kg/m², and the risk increased when BMI was below this value. This finding is consistent with the "obesity paradox"—individuals with mild overweight may gain survival advantages due to the secretion of cardioprotective adipokines by adipose tissue [32-34]. In addition, smoking history (including former and current smoking) and second-hand smoke exposure significantly increased mortality risk, emphasizing the importance of smoking cessation and reducing second-hand smoke exposure for improving treatment efficacy and survival [34]. These results suggest that chemotherapy regimens should be formulated based on individual patient characteristics; nutritional status should be closely monitored during treatment with targeted support provided; and research on sex-specific treatment differences needs to be strengthened.

The Prognostic Significance of Tumor Biological Characteristics

SCLC is often diagnosed at an advanced stage and is typically more aggressive. Tumors at T3/T4 stages (locally advanced or invasive tumors) are closely associated with increased tumor burden and local invasion, which may lead to chemotherapy resistance and elevated all-cause mortality risk [35]. The N stage reflects the spread of cancer cells to lymph nodes, leading to increased drug resistance. Patients with advanced metastatic stages (M2) frequently lose the opportunity for curative surgery, making platinum-based chemotherapy the main palliative treatment modality. However, metastatic disease, especially distant organ metastasis, is associated with inherent chemotherapy resistance, resulting in an unfavorable prognosis [36].

Potential Prognostic Value of CD56

The association between CD56 positivity and lung cancer prognosis remains controversial. Some studies have found a correlation with poor prognosis [37,38], while others have not observed a significant association [39]. This discrepancy may be related to sample size and patient population heterogeneity, and larger-scale multicenter studies are needed to clarify the relationship between the two.

Prognostic Value of PRO

This study suggested that the symptom burden reflected by the MDASI-LC score is related to survival time, with patients in the high-symptom burden group having a significantly increased mortality risk. This is consistent with previous research conclusions on the prognostic value of health-related quality of life [40]; incorporating MDASI-LC into routine follow-up can timely capture symptom changes and provide a basis for individualized prognostic management.

Clinical Application Value of Biomarkers

Blood biomarkers directly reflect tumor burden and biological characteristics. Elevated prechemotherapy biomarker levels have been associated with unfavorable chemotherapy outcomes in patients with lung cancer [41-44]. For patients with abnormal biomarkers, early interventions, such as anti-inflammatory and anti-infective treatments, can improve prognosis.

Limitations

This study has several limitations, which need to be taken into account when interpreting the results. The study used retrospective single-center data, which might limit the external validity and universality of the constructed model. Although the sample size of 1278 patients is relatively large for a single-center study, the complexity of the ML framework increases the risk of overfitting. We did not adjust the performance estimates using optimistic correction or penalty

methods, which might overestimate the model's predictive ability. The robustness of the feature attribution results was not verified under different model parameter settings, so the stability of these key prognostic thresholds still needs to be further confirmed.

Among the 1278 enrolled patients with lung cancer undergoing chemotherapy, the survival rates at 1-, 3-, and 5-year follow-ups were 31.9% (95% CI 29.4%-34.6%), 7.4% (95% CI 6.0%-9.0%), and 1.5% (95% CI 0.9%-2.4%), respectively. A smaller sample size may increase the variability of the AUC statistic. This might be part of the reason why the 5-year AUC (0.915) is higher than the 1-year (0.740) and 3-year (0.777) AUC, which needs to be taken into account when interpreting the long-term predictive performance of the model.

In this study, 84 models were constructed on the training set, and then, the C-index was used on the test set to determine the best model. This process may violate the principle that the test set should only be used once to evaluate the performance of the final, prespecified model, which may lead to an overestimation of the model's C-index. In future research, model selection and hyperparameter tuning can be attempted to be accomplished solely through cross-validation of the training set, with the test set reserved only for evaluating the prespecified final model performance.

Conclusions

This study developed an innovative ML model to predict all-cause mortality risk in patients with lung cancer receiving chemotherapy. The model demonstrates acceptable discrimination, accuracy, and clinical efficacy, offering clinicians an interpretable tool for mortality risk assessment. By enabling personalized risk stratification, this model can guide targeted postchemotherapy care for patients with high risk, ultimately supporting the development of individualized treatment plans.

Acknowledgments

The authors are sincerely grateful to Professor Charles S Cleeland and Professor Shelley X Wang of The University of Texas MD Anderson Cancer Center and Professor Qiuling Shi from Chongqing Medical University, who provided the MD Anderson Symptom Inventory for Lung Cancer questionnaire and other assistance for this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due patient privacy but are available from the corresponding author on reasonable request.

Funding

This study was supported by the National Natural Science Foundation of China Grant (82173609, 82373678, 82574198, 81673267), Tertiary Education Scientific Research Project of Guangzhou Municipal Education Bureau (202235407), Key Research Fund for colleges and universities of Guangdong Education Department (2023ZDZX2049), Guangzhou Science and Technology Plan Project (2025A03J3605), Guangzhou Science and Technology Plan Project (2025A03J4485), and Guangdong Provincial Medical Science and Technology Research Fund Project (A2023352).

Authors' Contributions

Conceptualization: J Zou

Data curation: AL

Validation: ZL (lead), BR (supporting)

Formal analysis: JH (lead), CX (supporting)

Investigation: J Zou (lead), J Zhang (supporting), LL (supporting), FL (supporting)

Methodology: JH (lead), CX (equal)

Project administration: JL (lead), YD (equal), JC (supporting), FQ (supporting), LY (supporting)

Funding acquisition: JL (lead), LY (supporting), JC (supporting), XZ (supporting)

Visualization: JH (lead), DX (equal)

Writing—original draft: JH (lead), J Zou (equal)

Writing—review and editing: KL (lead), JH (supporting), J Zou (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Complete baseline characteristics of chemotherapy-treated patients with lung cancer.

[\[DOCX File, 47 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

The evaluation of trajectory model fitting for the MD Anderson Symptom Inventory for Lung Cancer, Karnofsky Performance Status, and Zubrod Performance Score scales, the comparison of baseline data distribution before and after missForest interpolation, the results of univariate Cox regression, the comparison of baseline characteristics between the training and test sets, and the subgroup analysis of key values for feature attribution.

[\[DOCX File, 253 KB-Multimedia Appendix 2\]](#)

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229-263. [\[FREE Full text\]](#) [doi: [10.3322/caac.21834](https://doi.org/10.3322/caac.21834)] [Medline: [38572751](https://pubmed.ncbi.nlm.nih.gov/38572751/)]
2. Zhao Y, Mei T, Na F, Tian X, Ao R, Long X, et al. First-line treatment of driver gene-negative metastatic lung adenocarcinoma with malignant pleural effusion: should chemotherapy be combined with an immune checkpoint inhibitor or bevacizumab? *Invest New Drugs*. 2024;42(2):196-206. [\[FREE Full text\]](#) [doi: [10.1007/s10637-024-01424-4](https://doi.org/10.1007/s10637-024-01424-4)] [Medline: [38386170](https://pubmed.ncbi.nlm.nih.gov/38386170/)]
3. Wu K, House L, Liu W, Cho WCS. Personalized targeted therapy for lung cancer. *Int J Mol Sci*. 2012;13(9):11471-11496. [\[FREE Full text\]](#) [doi: [10.3390/ijms130911471](https://doi.org/10.3390/ijms130911471)] [Medline: [23109866](https://pubmed.ncbi.nlm.nih.gov/23109866/)]
4. Pei Q, Luo Y, Chen Y, Li J, Xie D, Ye T. Artificial intelligence in clinical applications for lung cancer: diagnosis, treatment and prognosis. *Clin Chem Lab Med*. 2022;60(12):1974-1983. [\[FREE Full text\]](#) [doi: [10.1515/cclm-2022-0291](https://doi.org/10.1515/cclm-2022-0291)] [Medline: [35771735](https://pubmed.ncbi.nlm.nih.gov/35771735/)]
5. Huang Z, Hu C, Chi C, Jiang Z, Tong Y, Zhao C. An artificial intelligence model for predicting 1-year survival of bone metastases in non-small-cell lung cancer patients based on XGBoost algorithm. *Biomed Res Int*. 2020;2020:3462363. [\[FREE Full text\]](#) [doi: [10.1155/2020/3462363](https://doi.org/10.1155/2020/3462363)] [Medline: [32685470](https://pubmed.ncbi.nlm.nih.gov/32685470/)]
6. Sibille L, Seifert R, Avramovic N, Vehren T, Spottiswoode B, Zuehlsdorff S, et al. 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*. 2020;294(2):445-452. [doi: [10.1148/radiol.2019191114](https://doi.org/10.1148/radiol.2019191114)] [Medline: [31821122](https://pubmed.ncbi.nlm.nih.gov/31821122/)]
7. Goldstraw P, Crowley J, Chansky K, Giroux DJ, Groome PA, Rami-Porta R, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J Thorac Oncol*. 2007;2(8):706-714. [\[FREE Full text\]](#) [doi: [10.1097/JTO.0b013e31812f3c1a](https://doi.org/10.1097/JTO.0b013e31812f3c1a)] [Medline: [17762336](https://pubmed.ncbi.nlm.nih.gov/17762336/)]
8. López-Encuentra A, López-Ríos F, Conde E, García-Luján R, Suárez-Gauthier A, Mañes N, et al. Composite anatomical-clinical-molecular prognostic model in non-small cell lung cancer. *Eur Respir J*. 2011;37(1):136-142. [\[FREE Full text\]](#) [doi: [10.1183/09031936.00028610](https://doi.org/10.1183/09031936.00028610)] [Medline: [20817702](https://pubmed.ncbi.nlm.nih.gov/20817702/)]
9. Zhang R, Chen C, Dong X, Shen S, Lai L, He J, et al. Independent validation of early-stage non-small cell lung cancer prognostic scores incorporating epigenetic and transcriptional biomarkers with gene-gene interactions and main effects. *Chest*. 2020;158(2):808-819. [\[FREE Full text\]](#) [doi: [10.1016/j.chest.2020.01.048](https://doi.org/10.1016/j.chest.2020.01.048)] [Medline: [32113923](https://pubmed.ncbi.nlm.nih.gov/32113923/)]
10. Chen YX, Chen J, Yin JY, Zhou HH, He BM, Liu ZQ. Non-coding RNA polymorphisms (rs2910164 and rs1333049) associated with prognosis of lung cancer under platinum-based chemotherapy. *Front Pharmacol*. 2021;12:709528. [\[FREE Full text\]](#) [doi: [10.3389/fphar.2021.709528](https://doi.org/10.3389/fphar.2021.709528)] [Medline: [34603024](https://pubmed.ncbi.nlm.nih.gov/34603024/)]

11. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet.* 2012;44(10):1111-1116. [FREE Full text] [doi: [10.1038/ng.2405](https://doi.org/10.1038/ng.2405)] [Medline: [22941189](https://pubmed.ncbi.nlm.nih.gov/22941189/)]
12. Cleeland CS, Mendoza TR, Wang XS, Chou C, Harle MT, Morrissey M, et al. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. *Cancer.* 2000;89(7):1634-1646. [doi: [10.1002/1097-0142\(20001001\)89:7<1634::aid-cnrc29>3.0.co;2-v](https://doi.org/10.1002/1097-0142(20001001)89:7<1634::aid-cnrc29>3.0.co;2-v)] [Medline: [11013380](https://pubmed.ncbi.nlm.nih.gov/11013380/)]
13. Wei X, Yu H, Dai W, Mu Y, Wang Y, Liao J, et al. Patient-reported outcomes of video-assisted thoracoscopic surgery versus thoracotomy for locally advanced lung cancer: a longitudinal cohort study. *Ann Surg Oncol.* 2021;28(13):8358-8371. [doi: [10.1245/s10434-021-09981-1](https://doi.org/10.1245/s10434-021-09981-1)] [Medline: [33880671](https://pubmed.ncbi.nlm.nih.gov/33880671/)]
14. Calvert M, Kyte D, Mercieca-Bebber R, Slade A, Chan A, King MT, et al. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO Extension. *JAMA.* 2018;319(5):483-494. [FREE Full text] [doi: [10.1001/jama.2017.21903](https://doi.org/10.1001/jama.2017.21903)] [Medline: [29411037](https://pubmed.ncbi.nlm.nih.gov/29411037/)]
15. Efficace F, Bottomley A, Smit E, Lianes P, Legrand C, Debruyne C, et al. Is a patient's self-reported health-related quality of life a prognostic factor for survival in non-small-cell lung cancer patients? A multivariate analysis of prognostic factors of EORTC study 08975. *Ann Oncol.* 2006;17(11):1698-1704. [FREE Full text] [doi: [10.1093/annonc/mdl183](https://doi.org/10.1093/annonc/mdl183)] [Medline: [16968876](https://pubmed.ncbi.nlm.nih.gov/16968876/)]
16. Deliu N, Cottone F, Collins GS, Anota A, Efficace F. Evaluating methodological quality of Prognostic models Including Patient-reported Health outcomes in oncology (EPIPHANY): a systematic review protocol. *BMJ Open.* 2018;8(10):e025054. [FREE Full text] [doi: [10.1136/bmjopen-2018-025054](https://doi.org/10.1136/bmjopen-2018-025054)] [Medline: [30361409](https://pubmed.ncbi.nlm.nih.gov/30361409/)]
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594. [FREE Full text] [doi: [10.1136/bmj.g7594](https://doi.org/10.1136/bmj.g7594)] [Medline: [25569120](https://pubmed.ncbi.nlm.nih.gov/25569120/)]
18. Mendoza TR, Wang XS, Lu C, Palos GR, Liao Z, Mobley GM, et al. Measuring the symptom burden of lung cancer: the validity and utility of the lung cancer module of the M. D. Anderson Symptom Inventory. *Oncologist.* 2011;16(2):217-227. [FREE Full text] [doi: [10.1634/theoncologist.2010-0193](https://doi.org/10.1634/theoncologist.2010-0193)] [Medline: [21285393](https://pubmed.ncbi.nlm.nih.gov/21285393/)]
19. Mor V, Laliberte L, Morris JN, Wiemann M. The Karnofsky Performance Status Scale. An examination of its reliability and validity in a research setting. *Cancer.* 1984;53(9):2002-2007. [doi: [10.1002/1097-0142\(19840501\)53:9<2002::aid-cnrc2820530933>3.0.co;2-w](https://doi.org/10.1002/1097-0142(19840501)53:9<2002::aid-cnrc2820530933>3.0.co;2-w)] [Medline: [6704925](https://pubmed.ncbi.nlm.nih.gov/6704925/)]
20. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982;5(6):649-655. [Medline: [7165009](https://pubmed.ncbi.nlm.nih.gov/7165009/)]
21. Nagin DS. Group-based trajectory modeling: an overview. *Ann Nutr Metab.* 2014;65(2-3):205-210. [doi: [10.1159/000360229](https://doi.org/10.1159/000360229)] [Medline: [25413659](https://pubmed.ncbi.nlm.nih.gov/25413659/)]
22. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-860. [doi: [10.1214/08-aos169](https://doi.org/10.1214/08-aos169)]
23. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* Oct 1, 2001;29(5):1189-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
24. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif Intell Med.* 2011;53(2):107-118. [doi: [10.1016/j.artmed.2011.06.006](https://doi.org/10.1016/j.artmed.2011.06.006)] [Medline: [21821401](https://pubmed.ncbi.nlm.nih.gov/21821401/)]
25. Zou H, Hastie T. Addendum: regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol.* 2005;67(5):768. [FREE Full text] [doi: [10.1111/j.1467-9868.2005.00527.x](https://doi.org/10.1111/j.1467-9868.2005.00527.x)]
26. Ianniello A, Sansovini M, Severi S, Nicolini S, Grana CM, Massri K, et al. Peptide receptor radionuclide therapy with (177)Lu-DOTATATE in advanced bronchial carcinoids: prognostic role of thyroid transcription factor 1 and (18)F-FDG PET. *Eur J Nucl Med Mol Imaging.* 2016;43(6):1040-1046. [doi: [10.1007/s00259-015-3262-8](https://doi.org/10.1007/s00259-015-3262-8)] [Medline: [26611427](https://pubmed.ncbi.nlm.nih.gov/26611427/)]
27. Li S, Wang W, Yu H, Zhang S, Bi W, Sun S, et al. Characterization of genomic instability-related genes predicts survival and therapeutic response in lung adenocarcinoma. *BMC Cancer.* 2023;23(1):1115. [FREE Full text] [doi: [10.1186/s12885-023-11580-0](https://doi.org/10.1186/s12885-023-11580-0)] [Medline: [37974107](https://pubmed.ncbi.nlm.nih.gov/37974107/)]
28. Feng J, Lan Y, Liu F, Yuan Y, Ge J, Wei S, et al. Combination of genomic instability score and TP53 status for prognosis prediction in lung adenocarcinoma. *NPJ Precis Oncol.* 2023;7(1):110. [FREE Full text] [doi: [10.1038/s41698-023-00465-x](https://doi.org/10.1038/s41698-023-00465-x)] [Medline: [37907595](https://pubmed.ncbi.nlm.nih.gov/37907595/)]
29. Wisnivesky JP, Halm EA. Sex differences in lung cancer survival: do tumors behave differently in elderly women? *J Clin Oncol.* 2007;25(13):1705-1712. [doi: [10.1200/JCO.2006.08.1455](https://doi.org/10.1200/JCO.2006.08.1455)] [Medline: [17470862](https://pubmed.ncbi.nlm.nih.gov/17470862/)]
30. Skuladottir H, Olsen JH. Can reproductive pattern explain better survival of women with lung cancer? *Acta Oncol.* 2006;45(1):47-53. [doi: [10.1080/02841860500374455](https://doi.org/10.1080/02841860500374455)] [Medline: [16464795](https://pubmed.ncbi.nlm.nih.gov/16464795/)]
31. Moore KA, Mery CM, Jaklitsch MT, Estocin AP, Bueno R, Swanson SJ, et al. Menopausal effects on presentation, treatment, and survival of women with non-small cell lung cancer. *Ann Thorac Surg.* 2003;76(6):1789-1795. [doi: [10.1016/s0003-4975\(03\)01024-5](https://doi.org/10.1016/s0003-4975(03)01024-5)] [Medline: [14667585](https://pubmed.ncbi.nlm.nih.gov/14667585/)]
32. Johnston EM, McIntyre LM, Hoff JA, Bepler G. The effect of orchiectomy on lung cancer survival. *Anticancer Res.* 1999;19(6C):5567-5570. [Medline: [10697618](https://pubmed.ncbi.nlm.nih.gov/10697618/)]

33. Antonopoulos AS, Tousoulis D. The molecular mechanisms of obesity paradox. *Cardiovasc Res*. 2017;113(9):1074-1086. [doi: [10.1093/cvr/cvx106](https://doi.org/10.1093/cvr/cvx106)] [Medline: [28549096](https://pubmed.ncbi.nlm.nih.gov/28549096/)]
34. Caan BJ, Cespedes Feliciano EM, Kroenke CH. The importance of body composition in explaining the overweight paradox in cancer-counterpoint. *Cancer Res*. 2018;78(8):1906-1912. [FREE Full text] [doi: [10.1158/0008-5472.CAN-17-3287](https://doi.org/10.1158/0008-5472.CAN-17-3287)] [Medline: [29654153](https://pubmed.ncbi.nlm.nih.gov/29654153/)]
35. Bergman M, Fountoukidis G, Smith D, Ahlgren J, Lambe M, Valachis A. Effect of smoking on treatment efficacy and toxicity in patients with cancer: a systematic review and meta-analysis. *Cancers (Basel)*. 2022;14(17):4117. [FREE Full text] [doi: [10.3390/cancers14174117](https://doi.org/10.3390/cancers14174117)] [Medline: [36077654](https://pubmed.ncbi.nlm.nih.gov/36077654/)]
36. Wu Y, Yuan M, Wang C, Chen Y, Zhang Y, Zhang J. T lymphocyte cell: a pivotal player in lung cancer. *Front Immunol*. 2023;14:1102778. [doi: [10.3389/fimmu.2023.1102778](https://doi.org/10.3389/fimmu.2023.1102778)] [Medline: [36776832](https://pubmed.ncbi.nlm.nih.gov/36776832/)]
37. Feng J, Sheng H, Zhu C, Qian X, Wan D, Su D, et al. Correlation of neuroendocrine features with prognosis of non-small cell lung cancer. *Oncotarget*. 2016;7(44):71727-71736. [FREE Full text] [doi: [10.18632/oncotarget.12327](https://doi.org/10.18632/oncotarget.12327)] [Medline: [27687592](https://pubmed.ncbi.nlm.nih.gov/27687592/)]
38. Wang X, Zhang Y, Hu M, Wang R, Liu L, Qian K, et al. Prognostic and predictive value of thyroid transcription factor-1, CD56, P40 and other clinical characteristics in small cell lung cancer patients. *Zhongguo Fei Ai Za Zhi*. 2017;20(8):522-527. [FREE Full text] [doi: [10.3779/j.issn.1009-3419.2017.08.04](https://doi.org/10.3779/j.issn.1009-3419.2017.08.04)] [Medline: [28855032](https://pubmed.ncbi.nlm.nih.gov/28855032/)]
39. Georgakopoulou V, Zygoris E, Damaskos C, Pierrakou A, Papalexis P, Garmpis N, et al. Prognostic value of the immunohistochemistry markers CD56, TTF-1, synaptophysin, CEA, EMA and NSE in surgically resected lung carcinoid tumors. *Mol Clin Oncol*. 2022;16(2):31. [FREE Full text] [doi: [10.3892/mco.2021.2464](https://doi.org/10.3892/mco.2021.2464)] [Medline: [34984102](https://pubmed.ncbi.nlm.nih.gov/34984102/)]
40. Wang XS, Shi Q, Lu C, Basch EM, Johnson VE, Mendoza TR, et al. Prognostic value of symptom burden for overall survival in patients receiving chemotherapy for advanced nonsmall cell lung cancer. *Cancer*. 2010;116(1):137-145. [FREE Full text] [doi: [10.1002/cncr.24703](https://doi.org/10.1002/cncr.24703)] [Medline: [19852033](https://pubmed.ncbi.nlm.nih.gov/19852033/)]
41. Barlési F, Gimenez C, Torre J, Doddoli C, Mancini J, Greillier L, et al. Prognostic value of combination of Cyfra 21-1, CEA and NSE in patients with advanced non-small cell lung cancer. *Respir Med*. 2004;98(4):357-362. [FREE Full text] [doi: [10.1016/j.rmed.2003.11.003](https://doi.org/10.1016/j.rmed.2003.11.003)] [Medline: [15080079](https://pubmed.ncbi.nlm.nih.gov/15080079/)]
42. Mandrekar SJ, Schild SE, Hillman SL, Allen KL, Marks RS, Mailliard JA, et al. A prognostic model for advanced stage nonsmall cell lung cancer. Pooled analysis of North Central Cancer Treatment Group trials. *Cancer*. 2006;107(4):781-792. [FREE Full text] [doi: [10.1002/cncr.22049](https://doi.org/10.1002/cncr.22049)] [Medline: [16847887](https://pubmed.ncbi.nlm.nih.gov/16847887/)]
43. Kanoh Y, Abe T, Masuda N, Akahoshi T. Progression of non-small cell lung cancer: diagnostic and prognostic utility of matrix metalloproteinase-2, C-reactive protein and serum amyloid A. *Oncol Rep*. 2013;29(2):469-473. [doi: [10.3892/or.2012.2123](https://doi.org/10.3892/or.2012.2123)] [Medline: [23135313](https://pubmed.ncbi.nlm.nih.gov/23135313/)]
44. Albain KS, Crowley JJ, LeBlanc M, Livingston RB. Survival determinants in extensive-stage non-small-cell lung cancer: the Southwest Oncology Group experience. *J Clin Oncol*. 1991;9(9):1618-1626. [doi: [10.1200/JCO.1991.9.9.1618](https://doi.org/10.1200/JCO.1991.9.9.1618)] [Medline: [1651993](https://pubmed.ncbi.nlm.nih.gov/1651993/)]

Abbreviations

AUC: area under the curve
CD56: cluster of differentiation 56
C-index: concordance index
Enet: elastic net
GBM: gradient boosting machine
GBTM: group-based trajectory modeling
KPS: Karnofsky Performance Status
MDASI-LC: MD Anderson Symptom Inventory for Lung Cancer
ML: machine learning
PRO: patient-reported outcome
RCS: restricted cubic spline
RSF: random survival forest
SCLC: small cell lung cancer
Survival-SVM: survival support vector machine
ZPS: Zubrod Performance Score

Edited by A Coristine, A Castonguay; submitted 10.Feb.2025; peer-reviewed by C Wu, W Fu, D Surian; comments to author 28.Apr.2025; accepted 18.Nov.2025; published 18.Dec.2025

Please cite as:

Zou J, Huang J, Lu K, Lin A, Xie C, Zhang J, Rao B, Li Z, Xie D, Lu L, Luo F, Chen J, Yang L, Qiu F, Zhang X, Deng Y, Lu J
A Machine Learning Approach to Predicting Mortality Risk in Chemotherapy-Treated Lung Cancer: Machine Learning Model Development and Validation

JMIR Med Inform 2025;13:e72424

URL: <https://medinform.jmir.org/2025/1/e72424>

doi: [10.2196/72424](https://doi.org/10.2196/72424)

PMID: [41411039](https://pubmed.ncbi.nlm.nih.gov/41411039/)

©Jianjun Zou, Jinyi Huang, Katie Lu, Ao Lin, Chen Xie, Jinrong Zhang, Boqi Rao, Zhi Li, Dongming Xie, Ling Lu, Feng Luo, Jinbin Chen, Lei Yang, Fuman Qiu, Xin Zhang, Yibin Deng, Jiachun Lu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org/>), 18.Dec.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.