Using Nomogram and Machine Learning Models to Predict Non-Small Cell Lung Cancer Prognosis

Haike Lei

Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital

Chun Liu

Digital Health China Technologies,Ltd.,Co. <https://orcid.org/0000-0003-2337-2264>

Zheng Xu

Digital Health China Technologies,Ltd.,Co.

Na Hong

Digital Health China Technologies,Ltd., Co.

Xiaosheng Li

Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital

Wei Zhou

Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital

Mengchun Gong

Digital Health China Technologies,Ltd.,Co

Guixue Wang

MOE Key Lab for Biorheological Science and Technology, State and Local Joint Engineering Laboratory for Vascular Implants, College of Bioengineering Chongqing University

Ying Wang

Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital

Hong Zhou

Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital

Yongzhong Wu (✉ tohongying@163.com)

Chongqing University Cancer Hospital <https://orcid.org/0000-0003-4626-1428>

Research

Keywords: Outcome prediction, Overall survival, Cox regression model, Machine learning, Non-small cell lung cancer

Posted Date: June 21st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-616177/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Patients with non-small cell lung cancer (NSCLC) often have a poor prognosis. Overall survival (OS) prediction through the early diagnosis of cancer has many benefits, such as allowing providers to design the best treatment plan for patients. In this study, we aimed to evaluate the prognostic factors in NSCLC patients, construct a nomogram, and develop machine learning models to predict the OS. We also conducted feature importance analysis to understand how relevant factors of NSCLC patients impact their OS.

Results

Multiple machine learning models were adopted in a retrospective cohort of patients from 2010 to 2015 in the Surveillance, Epidemiology, and End Results (SEER) database. Independent prognostic factors for NSCLC were determined using Cox proportional hazards regression analysis. We modeled OS and vital status as the outcomes and constructed and validated a nomogram to predict the OS of NSCLC. Furthermore, we applied logistic regression, random forest, XGBoost, decision tree, multilayer perceptron, and LightGBM to predict the patients' vital status. We tested the prediction ability of the models and evaluated their performances using accuracy, sensitivity, specificity, precision, and the area under the receiver operating characteristic curve. A total of 34,567 patients selected from the SEER database that met our criteria were included in this study. The nomogram visualized the OS prediction results of the Cox regression model. Among the classifiers, XGBoost had the best prediction performance, with an area under the curve of 0.733.

Conclusions

The results demonstrated that machine learning-based classifier models are capable of predicting the outcomes of patients with NSCLC. And Cox regression model-based nomogram interpreted the results well and supports potential medical applications.

Background

Lung cancer is the leading cause of death in cancer patients, and non-small cell lung cancer (NSCLC) accounts for 83% of all lung cancer cases, with an incidence rate of 40.60 per 100,000 and a 5-year survival rate of 22.1% [1]. Since NSCLC has a high incidence and poor prognosis, it is particularly important to determine its prognosis. Currently, clinicians usually determine prognosis based on surgical pathological staging, but this staging only takes into account the primary tumor, regional lymph node involvement, and distant metastasis, ignoring the role of other prognostic factors; this method has a poor predictive effect [2]. There is an urgent need for a clinical prognostic assessment system for patients with non-small cell lung cancer that is based on a large amount of data, with high reliability and good predictive effect.

The National Cancer Institute established the Surveillance, Epidemiology and End Results (SEER) database in 1973. It is recognized as one of the most authoritative sources of oncology patient follow-up data in the world and provides reliable data support for clinical research [3].

Machine learning is a suitable method to address this problem because algorithms can learn quickly from a large number of patients to produce more accurate predictions than those of any set of clinical experts [4]. Many studies have assessed the survival of lung cancer patients by analyzing large datasets, such as those in the SEER database, using machine learning techniques, including logistic regression and support vector machine [5], as well as methods based on integrated clustering [6]. An artificial neural network was used to predict the survival of patients with NSCLC, with an overall accuracy of 83% [7]. Parsimonious Bayes and decision trees have also been used to predict the survival time in lung cancer with 90% accuracy [8].

In this study, we used the SEER database to extract and build an analysis cohort from a large sample of NSCLC patients, and the least absolute shrinkage and selection operator (LASSO) regression was used to explore the factors that have a significant impact on patient prognosis. The Cox risk regression model was used to predict the 1-, 3-, and 5-year overall survival of patients and to present them in a nomogram. Multiple machine learning methods have also been used to construct NSCLC survival prediction models and compare their model parameters to select the best one for the decision support of clinicians during patient treatment.

Results

Dataset and Patient Inclusion

Anonymized patient data were obtained from the SEER database by the National Cancer Institute's SEER*Stat software [3]. The SEER program collected data on the incidence and survival from US regions, accounting for 28% of the national population [9]. We screened patients aged 15-64 years with the primary site diagnosed at the trachea, main bronchus, and lung between 2004 and 2015. We included patients with non-small cell lung cancer diagnosed through histology that had complete survival time and active follow-up data. Tumor stage was coded according to the American Joint Committee on Cancer TNM staging system, 6th edition [10]. Patients with cancer diagnosed only on autopsy or death certificates were excluded. To improve data completeness, we also excluded patients whose stage, grade, and race were missing. The total number of cases included was 34,567.

Feature Selection

Twenty-seven patient characteristics were selected from the SEER database. Because some variables were not in numeric format, for analysis and research purpose, they were assigned values according to Table S1. Some categorical variables, such as histological classification, race, surgery status, and

radiotherapy status, were processed by one-hot recoding. As the age of the patients was recorded as a range in the SEER database, the median value of the range was applied for this entry.

To simplify the model and reduce irrelevant variables, we used LASSO regression to screen 10 variables that are the most relevant to the outcome. The coefficient values are listed in Table 1. Among them, “regional nodes examined” and “regional nodes positive” have a strong correlation with N, and N is widely used in clinical practice. Therefore, we selected only N from the three. The remaining eight variables were used for the subsequent prediction and analysis.

Table 1. Variables selected by LASSO regression.

Variable	Mets at dx	N	Grade	Regional Nodes Examined	Sex
Coef value	0.131	0.109	0.104	0.082	0.081
Variable	Regional Nodes Positive	Stage	Extension	Age	Tumor Size
Coef value	0.072	0.070	0.064	0.061	0.041

Mets at dx: distant metastases (title of one parameter in the SEER database).

Coef: regression coefficient.

Multivariate Cox Regression Analysis

The multivariate Cox regression analysis showed that the 8 variables selected by LASSO regression, which were significantly related to the outcome, were all independent factors affecting the prognosis of NSCLC patients (all $P < 0.001$) (Table 2). Figure 1 shows the distribution of the number of patients and the hazard ratio (HR) and P values of the eight groups of variables.

The HR was used to evaluate the relative risk of a variable. A hazard ratio greater than 1 indicates that the variable is positively correlated with the probability of death and negatively correlated with survival time, and a hazard ratio of less than 1 indicates the opposite. For example, compared with female patients, male NSCLC patients have a higher probability of death, and the difference is significant.

Table 2. Multivariate Cox regression analysis on overall survival of non-small cell lung cancer patients.

Variables	HR	95% CI for HR		P
		Lower	Upper	
Age	1.0189	1.0163	1.0216	<0.001
Sex				
Female	1			—
Male	1.3341	1.2932	1.3763	<0.001
Grade	1.2632	1.2352	1.2918	<0.001
Stage				
Distant	1			—
Localized	0.7466	0.6863	0.8123	<0.001
Regional	0.8046	0.7517	0.8611	<0.001
N				
N0	1			—
N1	1.5325	1.4551	1.6140	<0.001
N2	2.1155	2.0200	2.2154	<0.001
N3	3.1735	2.9305	3.4366	<0.001
NX	3.0994	2.5198	3.8123	<0.001
Metastasis	1.0192	1.0174	1.0209	<0.001
Extension	1.0007	1.0006	1.0007	<0.001
Tumor Size	1.0008	1.0006	1.0010	<0.001

CI: confidence interval

Construction and Verification of Prognostic Prediction Nomogram

The nomogram was based on the Cox regression risk ratio model constructed before and displayed the overall survival rates of patients at 1, 3, and 5 years in a graphical style.

Figure 2 shows the points corresponding to different values of the variables. We can calculate the patient's total points by adding and then create a vertical line to obtain the patient's 1-, 3-, and 5- survival probabilities. The C-index for the prediction model was 0.709 (SE = 0.002).

For the C-index, a value greater than 0.7 usually indicates a relatively good distinction, and the closer the C-index is to 1, the better the prediction accuracy. The calibration curve was used to compare the consistency between the prediction result of the nomogram and the result of the Kaplan-Meier method. If the two results are identical, the calibration curve coincides with the diagonal. In this study, calibration curves for patients undergoing NSCLC surgery at 1, 3, and 5 years were plotted. As shown in Figure S1, the calibration curve is relatively close to the diagonal, indicating that the prediction results of the nomogram model are highly consistent with the prediction results of the Kaplan-Meier method.

Evaluation Results

Several machine learning methods, including logistic regression, random forest, XGBoost, decision tree, and LightGBM, were selected to build the prognostic models. The prediction accuracy and model performance evaluation of the five machine learning algorithm models are compared in Table 3 and Figure 3.

Table 3. Model performance on the SEER database

Model	Accuracy	Sensitivity	Specificity	Precision	AUC
Logistic Regression	0.657	0.539	0.761	0.666	0.716
Random Forest	0.669	0.583	0.745	0.669	0.726
XGBoost	0.673	0.584	0.752	0.676	0.733
Decision Tree	0.666	0.581	0.741	0.666	0.717
LightGBM	0.668	0.637	0.695	0.649	0.733

AUC: area under the receiver operating characteristic (ROC)

Overall, the XGBoost model presented the best classification results, with AUC of 0.733, among all five classifiers (Table 3). The LightGBM model showed equivalent AUC as XGBoost and a more balanced performance (sensitivity = 0.637, specificity = 0.695) when comparing the difference between sensitivity and specificity. The classification model with the worst performance was logistic regression, with AUC of 0.716.

Discussion

Multivariate Cox regression analysis ($P<0.001$) demonstrated that eight variables (age, gender, grade, stage, N, metastasis, extension, and tumor size) were related to the final outcome of NSCLC patients (Table 2 and Figure 1). The results show that the older the age and the lower the degree of tumor differentiation, the higher the overall survival rate of NSCLC patients. The degree of involvement of the distal lymph nodes has a greater impact on the survival of patients due to its outstanding absolute HR. In addition, the prognosis of male patients is worse than that of female patients.

As a new statistical prediction model [11], the nomogram has many advantages over traditional prognostic models, such as high accuracy, flexibility, and ease of generalization, and it has been widely used with various tumor types, including liver, bladder, prostate, cervical, and gastric cancers [12-16]. In this study, a nomogram was constructed to predict the 1-, 3-, and 5-year OS of NSCLC patients using eight variables that showed statistical significance in a multifactorial Cox risk regression analysis. The accuracy and reliability of the model were evaluated using the C-index and calibration curve analysis. C-index, which ranges from 0.5 to 1, where 1 means complete discrimination and 0.5 means no discrimination, was used for the measurement of the nomogram between performance and predicted results. The C-index of the nomogram for OS in our study was 0.709, which shows that the model has the capability of prediction. For the calibration curve, the horizontal coordinate of the graph is the predicted probability, and the probability of the event occurring is predicted using the prediction model, with 0 to 1 indicating that the probability of the event occurring is 0-100%. The vertical coordinates represent the actual incidence of the patient. Blue, red, and green are fitted lines for 1-, 3-, and 5-year OS, respectively. The fitted line (colored) completely overlaps the reference line (black) if the predicted value is equal to the actual value for each probability; if the predicted value is greater than the actual value, that is, the risk is overestimated, the fitted line is below the reference line. In this study, all three fitted lines overlapped well with the reference line, indicating the high accuracy of the prediction model. The 3- and 5-year OS have the best prediction at 60-70%, while 1-year OS has the best prediction at around 95%.

Logistic regression, random forest, XGBoost, decision tree, and LightGBM were selected to build the prognostic models (Table 3 and Figure 3). Among them, XGBoost obtained the highest accuracy, precision, and AUC, and is considered the best-performing model. LightGBM had the highest sensitivity, and the logistic regression model had the highest specificity. We also experimented with a deep learning model, multilayer perceptron, which underperformed with accuracy, sensitivity, specificity, precision, and AUC of 0.666, 0.598, 0.726, 0.659, and 0.723, respectively.

Cox regression is a time-dependent model, whereas the machine learning models used in this study are all classifiers. Only one of them is selected in most studies to analyze and predict patient survival. In this study, we not only compare the prediction advantages and disadvantages of the five classifiers on the vital status of NSCLC patients but also visualize the results of Cox regression models for the prediction of patients' 1-, 3-, and 5-year OS with a nomogram. Owing to the data, there is still much space for improvement in the final prediction results for both the Cox regression model and the best-performing machine learning model.

This study has limitations. We did not include all tumor prognostic factors in this preliminary study, so the number of indicators selected for modeling may not cover all the prognostic factors. In our future study, we will apply this model to stratified analysis on Asia patients from SEER database and chose Asia local hospital datasets for comparison, and explore the characteristic variables that have the greatest impact on survival outcomes on Asia patients and verify the model in real clinical settings.

Conclusions

In this study, we took preprocessed data on NSCLC patients derived from the SEER database, screened the characteristics by LASSO regression methods, and selected 8 sets of independent variables that had the greatest impact on survival outcomes among 27 sets of characteristic variables. We then used the Cox regression model to analyze the eight groups of variables of patients and produce nomograms to predict the 1-, 3-, and 5-year OS of patients. The C-index and calibration curves showed that they all had a good predictive effect. We also compared the prediction results of the five machine learning models for patient survival outcomes, and XGBoost achieved the best performance. We believe these results are useful for predicting patient outcomes and potentially help improve clinical decision making.

Methods

Analysis Methods

Feature selection method

The LASSO method was applied to select the most relevant features for the outcome. Here, the top ten features were chosen from the twenty-seven extracted features. Furthermore, since there were three strongly correlated variables among the top ten features, we chose the most commonly used one and discarded the other two. Finally, eight features were utilized in both machine learning models and nomograms.

Multivariate Cox regression analysis

Multivariate Cox regression analysis was applied to the variables obtained by LASSO regression to determine whether they were all independent factors. A nomogram based on Cox regression model was performed to predict NSCLC patients on 1-, 3-, and 5 years overall survival (OS).

Construction and verification of prognostic prediction nomogram

We used R software (version 4.0.4) to construct a survival prediction model for patients with non-small cell lung cancer surgery based on Cox proportional hazards. The concordance-index (C-index) was used to test the accuracy of the nomogram prediction, and a calibration curve was used to evaluate the predictive ability of the nomogram. The self-sampling and replacement method (bootstrapping) was used to verify the model internally to reduce the over-fitting of the model. The program packages "rccorp.cens" and "Himsc" in R language were used in this study.

Machine learning methods

All features were normalized before putting them into the classification models. Using survival status (dead/alive) as the predictor class, the classification model incorporated the following methods: logistic regression, random forest, XGBoost, decision tree, multilayer perceptron, and LightGBM. Because the

dataset (dead: 16,230, alive: 18,337) was not highly biased, it was unnecessary to apply any sampling method to adjust the balance of the dataset. The data were randomly split into the training and testing datasets with 80% and 20% of all patients, respectively. All analyses were performed using Python 3.6.

Evaluation Methods

The classification models were assessed using the area under the receiver operating characteristic (AUC) curve, sensitivity (also known as recall), specificity, accuracy, and precision. The foundation of these assessment variables comes from the four possible outcomes (TP = true positives, TN = true negatives, FP = false positives, FN = false negatives) of the binary classifier. Accuracy accounts for the fraction of correct predictions for both true positive and true negative among all subjects; which is suitable for a balanced dataset. Computed by plotting sensitivity as a function of 1-specificity, the area under the receiver operating characteristic curve is widely used as a performance measurement for classification problems at various threshold settings. A higher AUC value indicates a better model that distinguishes between classes. In this study, false positives (e.g., survivor is predicted as non-survivor) may be treated with overmedication, while false negatives (e.g., non-survivors are predicted to be survivors) will not take any extra actions for early prevention. Both cases should be avoided. Precision and recall are good measures to determine when the costs of false positives and false negatives are too high.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Abbreviations

AUC: area under the curve,

HR: hazard ratio,

LASSO: least absolute shrinkage and selection operator,

NSCLC: non-small cell lung cancer,

OS: overall survival,

ROC: receiver operating characteristic,

SEER: Surveillance, Epidemiology, and End Results.

Declarations

Ethics approval and consent to participate

The SEER database is a public database from which sensitive patient information has been removed. The data can only be used after it is submitted for research purposes and passed human review.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the SEER research data (18 Registries, 2019), Data application link <https://seer.cancer.gov/>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Chongqing Technology Innovation and Application Development Project (No. cstc2019jscx-fxydX0008), Chongqing Performance Incentive and Guidance Project for Scientific Research Institutions

(cstc2020jxjl130016), Chongqing Key Disease Prevention and Control

Technology Project (2019ZX002), Novel Coronavirus Infection and Prevention Emergency Scientific Research Special Project of Chongqing Municipal Education Commission, China (CQEO [2020] no.13)

Author Contributions

H. L, H. Z and Y. W were responsible for study design and conception; C. L and N. H collected the data, C. L, Z. X, X. L and W. Z were responsible for data processing and data analysis; M. G, G. W and Y. W interpreted the results; All authors drafted the manuscript. All authors revised the manuscript for important intellectual content.

Acknowledgements

Not applicable.

References

1. **SEER Cancer Statistics Review (CSR) 1975-2013 [R/OL]. [2016-09-20].** *National Cancer Institute*:http://seer.cancer.gov/csr/1975_2013/sections.html.
2. Ettinger DS, Wood DE, Akerley W, Bazhenova LA, Borghaei H, Camidge DR, Cheney RT, Chirieac LR, D'Amico TA, Dilling TJ *et al*: **NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 4.2016.** *Journal of the National Comprehensive Cancer Network : JNCCN* 2016, **14**(3):255-264.
3. **Surveillance, Epidemiology, and End Results (SEER).** <https://seer.cancer.gov/>. *National Cancer Institute* 2021.
4. Bartholomai JA, and Hermann B. Frieboes. : **Lung cancer survival prediction via machine learning regression, classification, and statistical techniques.** *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* 2018, **IEEE**.
5. Dmitriy Fradkin DSaIM: **Machine learning methods in the analysis of lung cancer survival data.** *DIMACS Technical Report* 2006, **2005-35**.
6. Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X: **Developing prognostic systems of cancer patients by ensemble clustering.** *Journal of biomedicine & biotechnology* 2009, **2009**:632786.
7. Chen YC, Ke WC, Chiu HW: **Risk classification of cancer survival using ANN with gene expression data from multiple laboratories.** *Computers in biology and medicine* 2014, **48**:1-7.
8. Jim GDaJAAaCM: **Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability.** *Journal of Computing* 2012, **4**:1-9.
9. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF: **Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population.** *Medical care* 2002, **40**(8 Suppl):lv-3-18.
10. Green FL AI, Apostolikas N: **American Joint Committee on Cancer Cancer Staging Manual 6th edn.** *Springer: New York, NY, USA* 2002.

11. Li G, Tian ML, Bing YT, Wang HY, Yuan CH, Xiu DR: **Nomograms predict survival outcomes for distant metastatic pancreatic neuroendocrine tumor: A population based STROBE compliant study.** *Medicine* 2020, **99**(13):e19593.
12. Wang YY, Xiang BD, Ma L, Zhong JH, Ye JZ, Wang K, Xing BC, Li LQ: **Development and Validation of a Nomogram to Preoperatively Estimate Post-hepatectomy Liver Dysfunction Risk and Long-term Survival in Patients With Hepatocellular Carcinoma.** *Annals of surgery* 2020.
13. Benoit L, Balaya V, Guani B, Bresset A, Magaud L, Bonsang-Kitzis H, Ngô C, Mathevet P, Lécuru F: **Nomogram Predicting the Likelihood of Parametrial Involvement in Early-Stage Cervical Cancer: Avoiding Unjustified Radical Hysterectomies.** *Journal of clinical medicine* 2020, **9**(7).
14. Yang Z, Bai Y, Liu M, Hu X, Han P: **Development and Validation of Prognostic Nomograms to Predict Overall and Cancer-Specific Survival for Patients with Adenocarcinoma of the Urinary Bladder: A Population-Based Study.** *Journal of investigative surgery : the official journal of the Academy of Surgical Research* 2020:1-8.
15. Kyei MY, Adusei B, Klufio GO, Mensah JE, Gepi-Attee S, Asante E: **Treatment of localized prostate cancer and use of nomograms among urologists in the West Africa sub-region.** *The Pan African medical journal* 2020, **36**:251.
16. Dong D, Tang L, Li ZY, Fang MJ, Gao JB, Shan XH, Ying XJ, Sun YS, Fu J, Wang XX *et al*: **Development and validation of an individualized nomogram to identify occult peritoneal metastasis in patients with advanced gastric cancer.** *Annals of oncology : official journal of the European Society for Medical Oncology* 2019, **30**(3):431-438.

Figures

Figure 1

Multivariate Cox regression analysis of overall survival of non-small cell lung cancer patients.

Figure 2

The nomogram of 1, 3, and 5-year overall survival prediction for non-small cell lung cancer patients by Cox risk regression model.

Figure 3

Area under the receiver operating characteristic (ROC) curve for different machine learning models.