Machine Learning Models for Risk Prediction of Lymph Nodes Metastasis in Non-Small Cell Lung Cancer: Development and Validation Study

Miaochun Cai
Southern Medical University
Dong Shen
Southern Medical University
Zhihao Li
Southern Medical University
Jianmeng Zhou
Southern Medical University
Yingjun Chen
Southern Medical University
Dan Liu
Southern Medical University
Yujie Zhang
Southern Medical University
Hong Shen
shenhong@smu.edu.cn
Chen Mao (maochen9@smu.edu.cn)
Southern Medical University

Research

Keywords: non-small cell lung cancer; lymph node metastasis; prediction model; machine learning

Posted Date: November 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-111012/v1

Abstract

Background: To develop and validate machine learning models for risk prediction of lymph node metastasis (LNM) in non-small cell lung cancer (NSCLC) using clinicopathologic parameters and immunohistochemical features.

Methods: From January 2010 to December 2019, 639 patients' data were continuously collected in Nanfang Hospital. We exacted immunohistochemical features and clinicopathological features from the electronic medical records of patients. We established two models (a full model and a selection model) and implemented three algorithms (random forest, support vector machine and penalized logistic regression). The model performance was evaluated in terms of discrimination (receiver operating characteristic curve (AUC)), calibration, and decision curve analysis.

Results: AUROC (area under receiver operating characteristic curve) analysis (also calibration curves) showed that the selection model (AUC values for training and testing, 0.843 and 0.840 respectively) and the full model constructed using random forest (AUC values for training and testing, 0.855 and 0.863 respectively) performed best among all models. Decision curve analysis depicted that the full model and the selection model using random forest was clinically useful. The model performance of the full model and the selection model were comparable.

Conclusion: The random forest model using clinicopathologic- immunohistochemical features can predict the LNM of NSCLC patients.

Background

Lung cancer is the most common cancer globally and the leading cause of cancer death globally [1]. Non-small cell lung cancer (NSCLC) accounts for 80% of all lung cancers [2]. Accurate identification of lymph node metastasis (LNM) is crucial in determining lung cancer patients' treatment. Therefore, the correct assessment of lymph node status is essential for lung cancer patients [3, 4]. Computed tomography (CT) is currently the most widely used technique for preoperative evaluation of LNM patients in clinical practice, but with low sensitivity and specificity [5–10] as CT data is based on the lymph node size to distinguish LNM from Non-LNM [11, 12]. Thus, preoperative identification of LNM is difficult, as smaller lymph nodes cannot be detected by imaging examination. Clinical and pathological features are commonly used predictors showing that objective and quantitative clinicopathological features can be used as prognostic or predictive features for LNM patients [13–17]. However, pathological diagnosis conclusions are often limited by sample sampling, and the ability to judge small lesions is low, which is likely to cause missed diagnosis. Clinical pathological specimens are often subjected to a large number of immunohistochemical examinations that is a fast, low-cost, and straightforward method.

A comprehensive analysis of a panel of factors, rather than single analysis, is the most powerful method to change clinical management [18]. In addition, machine learning has emerged as highly powerful tool for prediction model. Therefore, the main purpose of this study was to explore whether machine learning algorithms can make full use of the large amount of information provided by immunohistochemistry

combined with clinical-pathological characteristics to achieve a preliminary prediction of the risk of NSCLC patients with LNM. Develop and validate models that can accurately predict the presence of LNM, prevent lymph node-negative NSCLC patients from removing pathological lymph nodes, and remove lymph node-positive patients in time to achieve a good prognosis. We present the following article in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting checklist [19].

Methods

Patients

A total of 639 consecutive patients for lung cancer at Nanfang Hospital (Guangzhou, Guangdong, China) were identified retrospectively between January 2010 and December 2019. The medical records and histopathology reports were retrospectively analyzed. Inclusion criteria considered the conditions that (1) Patients with primary NSCLC underwent radical resection with routine systematic lymph nodes dissection. (2) The post-surgical diagnosis was NSCLC with the absence or presence of LNM. The exclusion criteria included (1) Patients with non-NSCLC such as mixed carcinoma, polymorphic carcinoma, atypical carcinoid, and mucoepidermoid carcinoma. (2) Radiotherapy or chemotherapy before surgery. (3) Patients underwent an incomplete resection or were without mediastinal node dissection. (4) Missing key information. Finally, 152 eligible patients were enrolled. The inclusion and exclusion criteria process is shown in Fig. 1. The institutional review board at Nanfang Hospital approved the ethical approval and the informed consent requirement was waived.

Predictors

Based on literature reviews [15, 18, 20–24], clinicians and pathologists' opinions on factors that may increase the risk of lymph node metastasis combined with our existing data, we identified 17 potential predictors including patients of basic information, clinicopathologic parameters, and immunohistochemical features. Except for the maximum tumor diameter, which is a continuous variable, the others are categorical variables. The definition of predictors is shown in Appendix A1.

Outcome

The outcome is lymph node status for NSCLC patients. All the patients underwent anatomical lung resection and systematic nodal dissection by thoracic surgeons. Experienced pulmonary pathologists histologically assessed all resected tumor specimens and nodal samples, and a final diagnosis was evaluated based on the WHO classification. The medical records and histopathology reports were retrospectively analyzed.

Model Development And Validation

We randomly divided the complete original dataset into 80% as the training set and 20% as the validation set. The analysis process consists of four main stages: data preprocessing, feature selection, construction of prediction models, and model evaluation. The first stage includes: data preprocessing incorporates two items of the data standardization, processing of missing values. As long as one categorical variable data is missing, we delete the row of data. For continuous variables, random forest imputation was used. Secondly, in the selection models, least absolute shrinkage and selection operator (LASSO) algorithms under 10-fold cross-validation were used to select features related to the outcome models. In the full models, we chose all features to build the models. In the third stage, based on the selected features, we conducted three algorithms included random forest, support vector machine, and penalized logistic regression to build prediction models. Finally, we used the area under the receiver operating characteristic (AUC), accuracy, sensitivity, specificity, calibration curves and decision curve analysis (DCA) to evaluate models. The calibration curve is usually used to evaluate the consistency or the degree of calibration, that is, the difference between the predicted value and the true value. DCA is a novel method of assessing the clinical prediction model used to help identify high-risk patients for intervention and low-risk patients to avoid over-treatment.

Statistical Analysis

Statistical analysis was conducted with R software (version 4.0.2; http://www.Rproject.org) and Python (version 3.7.0; https://www.python.org). The models were programmed in Python using the sklearn library. The AUROC curves, calibration curves and DCA were generated using the “ggplot2” package, “rms" package and "dca" package respectively. See the appendix A2 for the adjusted parameters of each algorithm. The Chi-square test was tested to compare the two counting data sets and two independent sample t-test for measurement data. Statistical analyses were all two-sided, with the p-value set at .05 while the p-value was set at. 01 in the Delong test. The AUC, accuracy, sensitivity and specificity were used for testing discrimination. The calibration curves were performed for testing calibration. The DCA was used to test clinical use.

Result

Characteristics of patients in cohorts

A total of 152 patients comprise the primary cohort. Patient characteristics in the training and validation cohort are given in Table 1. There were no significant differences between the two cohorts in lymph node prevalence (P = .830). The rate of LNM was 55.9% and 58.1% in the training and validation cohorts, respectively. Despite the temporary disconnect, no evidence for a statistically significant difference in demographic and clinical characteristics between the training cohort and the validation cohort, which confirmed training and validation cohort was reasonable.

Table 1
Characteristics of Patients in the Train and Validation Cohorts

Characteristic	Train cohort(n = 122)		*P*	Validation cohorts (n = 30)		*P*
	LNM(+)	LNM(-)		LNM(+)	LNM(-)	
Age, years			.304			.686
< 60	27(35.1)	20(44.4)		7(38.9)	6(46.2)	
≥ 60	50(64.9)	25(55.6)		11(61.1)	7(53.8)	
Sex			.886			.880
Female	23(29.9)	14(31.1)		6(33.3)	4(30.8)	
Male	54(70.1)	31(68.9)		12(66.7)	9(69.2)	
Lung diseases			.877			.524
No	52(67.5)	31(68.9)		9 (50.0)	5(38.5)	
Yes	25(32.5)	14(31.1)		9(50.0)	8(61.5)	
Hypertension			.106			.924
No	67 (87.0)	34 (75.6)		15(83.3)	11(84.6)	
Yes	10(13.0)	11(24.4)		3(16.7)	2(15.4)	
Diabetes			.370			.726
No	72(93.5)	40(88.9)		16 (88.9)	11 (84.6)	
Yes	5(6.5)	5(11.1)		2(11.1)	2(15.4)	
Cardiovascular disease			.201			.260
No	61(79.2)	31(68.9)		12(66.7)	11 (84.6)	
Yes	16(20.8)	14(31.1)		6(35.3)	2(15.4)	
History of cancer			.121			-
No	75(97.4)	41(91.1)		18 (6.7)	13 (10.0)	
Yes	2(2.6)	4(8.9)		0(0.0)	0(0.0)	
Histological type			.292			.077
Adenocarcinoma	57(74.0)	32(71.1)		10(55.6)	7(53.8)	
Squamous cell carcinoma	18(23.4)	9(20.0)		8(44.4)	3(23.1)	
Other †	2(2.6)	4(8.9)		0(0.0)	3(23.1)	
Differentiation			< .001*			.178

Characteristic	Train cohort(n = 122)		*P*	Validation cohorts (n = 30)		*P*
	LNM(+)	LNM(-)		LNM(+)	LNM(-)	
Low	19(23.0)	9(37.5)		7(38.9)	9(69.2)	
Medium	31(43.2)	9(37.5)		9(50.0)	4(30.8)	
High	25(33.8)	6 (25.0)		2(11.1)	0(0.0)	
Distant metastasis			< .001*			.002*
No	68(88.3)	23(51.1)		17(94.4)	6(46.2)	
Yes	9(11.7)	22(48.9)		1 (5.6)	7(53.8)	
Tumor location			.385			.156
LUL	20(26.0)	9(20.0)		5(27.8)	0(0.0)	
LLL	5(6.5)	7(15.6)		5(27.8)	3(23.1)	
RUL	26(33.8)	15(33.3)		5(27.8)	4(30.8)	
RML	8(10.4)	5(11.1)		0(0.0)	2(15.4)	
RLL	15(19.5)	5(11.1)		3(16.7)	3(23.1)	
Others ‡	3(3.9)	4(8.9)		0(0.0)	1(7.7)	

Table 1
(continued)

Characteristic	Train cohort(n = 122)		*P*	Validation cohorts (n = 30)		*P*
	LNM(+)	LNM(-)		LNM(+)	LNM(-)	
Maximum diameter, mean(SD), cm	3.06(1.86)	2.79(1.77)	.422	5.00(2.25)	2.48(1.85)	.002*
CK5/6			.002*			.085
negative	75(97.4)	36(80.0)		18(100.0)	11(84.6)	
positive	2(2.6)	9(20.0)		0(0.0)	2(15.4)	
CK7			< .001*			.074
negative	64(83.1)	23(51.1)		15(83.3)	7(53.8)	
positive	13(16.9)	22 (48.9)		3(16.7)	6(46.2)	
TTF-1			.013*			.172
negative	66(85.7)	30(66.7)		16(88.9)	9(69.2)	
positive	11(14.3)	15(33.3)		2(11.1)	4(30.8)	
Napsin-A			.011*			.361
negative	72(93.5)	35(77.8)		17(94.4)	11(84.6)	
positive	5(6.5)	10(22.2)		1(5.6)	2(15.4)	
P63			.669			.811
negative	74(96.1)	42(93.3)		17(94.4)	12(92.3)	
positive	3(3.9)	3(6.7)		1(5.9)	1(7.7)	
NOTE. †: For example, squamous carcinoma or large cell carcinoma. ‡: the tumor with more than two different locations.						
Abbreviations: LNM, lymph node metastasis; SD, standard deviation; LUL, left upper lobe; LLL, left lower Lobe; RUL, right upper lobe; RML, right middle lobe; RLL, right lower Lobe;						
*P value < .05.						

Feature Selection

A total of 17 features were collected from the clinicopathologic-immunohistochemical features of the cohort. The 17 features were used to construct the full models. We selected five non-zero coefficients (distant metastasis, differentiation, CK7, hypertension, CK5/6) from 17 features as potential predictors using LASSO regression within the training cohort to build selection models (Appendix A3).

Discrimination

The AUC value, accuracy, sensitivity, specificity for each model is shown in Table 2 and AUROC curves are shown for each model in Fig. 2. There was good discrimination in both the training and the validation cohorts of each model. The results showed that the random forest was outstanding compared with other algorithms in the validation cohort, whether in the full or selection models. The random forest algorithm of the full model obtained the top AUC of 0.863(95%CI: 0.734–0.992), the top accuracy of 0.774 (95% CI: 0.589, 0.904), and the top sensitivity of 0.889 (95%CI: 0.639, 0.981) in the validation cohort. However, the selection model's random forest algorithm obtained the top accuracy of 0.774 (95%CI: 0.589, 0.904) and the top specificity of 0.846 (95%CI: 0.537, 0.973) in the validation cohort. The discrimination of the full model and the selection model was comparable. The Delong test showed that there was no statistical difference between the AUC value of the full model and the selected model in the random forest algorithm ($p > 0.05$) in the validation cohort. It shows that the full model and the selection model in the random forest algorithm are comparable.

Table 2
Classification performance for each model

Algorithms	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Train cohort				
RF (Selection)	0.843(0.767,0.920)	0.812(0.731,0.877)	0.870(0.770,0.933)	0.711(0.554,0.832)
RF (Full)	0.855(0.783–0.927)	0.812(0.731,0.877)	0.935(0.848,0.976)	0.600(0.444,0.739)
LR (Selection)	0.861(0.786–0.936)	0.803(0.722,0.870)	0.883(0.785,0.942)	0.667(0.509,0.796)
LR (Full)	0.840(0.762–0.918)	0.787(0.704,0.856)	0.870(0.770,0.933)	0.644(0.487,0.777)
SVM(Selection)	0.861(0.786–0.936)	0.836(0.758,0.897)	0.935(0.848,0.976)	0.667(0.509,0.796)
SVM (Full)	0.832(0.751–0.912)	0.746(0.659,0.820)	0.831(0.725,0.904)	0.600(0.443,0.739)
Validation cohort				
RF (Selection)	0.840(0.688–0.992)	0.774(0.589,0.904)	0.722(0.464,0.893)	0.846(0.537,0.973)
RF (Full)	0.863(0.734–0.992)	0.774(0.589,0.904)	0.889(0.639,0.981)	0.615(0.323,0.849)
LR (Selection)	0.812(0.658–0.966)	0.710(0.520,0.858)	0.778(0.519,0.926)	0.616(0.323,0.849)
LR (Full)	0.833(0.691–0.976)	0.710(0.520,0.858)	0.778(0.519,0.926)	0.615(0.323,0.849)
SVM(Selection)	0.832(0.751–0.912)	0.746(0.659,0.820)	0.831(0.725,0.904)	0.600(0.443,0.739)
SVM (Full)	0.825(0.659–0.990)	0.742(0.554,0.881)	0.833(0.577,0.956)	0.615(0.323,0.849)
Note: Selection, the selection models. Full, the full models.				
Abbreviations: AUC: the Area under the Receiver Operating Characteristic; RF: random forest; LR: logistic regression; SVM: support vector machine. CI: confidence interval.				

Model Calibration And Decision Curve Analysis

The calibration curve of six models in the validation cohort is shown in Fig. 3. The random forest's calibration curve for the full model offers the best agreement between the predictions and observations (Fig. 3D). The decision curve for each model is presented in Fig. 4. The decision curves displayed that the threshold

probability for the full model's random forest was between 16% -81%, while the random forest of the selection model is > 9%. Within these ranges, we use the models to predict LNM added more net benefit than treat all patients or treat none. Besides, within a larger threshold range, the selection model's net benefit was higher than the net benefit of the full model. Assuming that we choose a prediction probability of 60% to diagnose LNM and perform treatment, then for every 100 patients who use the selection model, 25 people can benefit from it without harming the interests of anyone else; for every 100 patients who use the full model, Only 14 people can benefit from it without harming anyone else's interests.

Discussion

In this study, we developed and validated machine learning models based on clinicopathologic and immunohistochemical parmeters to predict LNM patients with NSCLC. Our results showed that immunohistochemical features could provide more information to distinguish the absent or present LNM in patients with NSCLC. The combined determination with clinicopathological features demonstrated high discrimination, calibration, and clinical use to diagnose LNM.

For the construction of clinicopathological-immune features, the correlation between the predictor and the result was tested using the LASSO method to shrink the regression coefficient. The 17 candidate features were reduced to 5 potential predictors. LASSO is a statistical method that imposes L1 or L2 penalties to improve prediction accuracy and model interpretation, thereby generating sparse models [25]. In recent studies [26–28], LASSO regression has been used to integrate a single biomarker into a comprehensive analysis of multiple markers, and this method has been gradually developed and validated. To compare the feature screening effect of lasso regression, we also performed a full model incorporating all features. In this study, we selected three machine-learning algorithms to develop and validate models. It turned out that random forest is the best classifier on the entire dataset. These results were similar to other findings [29–31]. The two models' performance is comparable in the random forest algorithm, whether it is a full model or a selection model. Given that performance was comparable in the two random forest models, we are more inclined to choose the selection model of the random forest algorithm as the optimal model to obtain the greatest clinical utility with the fewest features. The selection model of the random forest algorithm showed good degree of discrimination (AUC, 0.84) and calibration in the training cohort, and then performed well in the validation cohort (AUC, 0.84). Given that the positive rates of LNM in the two cohorts are comparable, good discrimination means that the model is robust to prediction.

The evaluation of model performance is generally only analyzed from the degree of discrimination and calibration. However, AUROC only considers the specificity and sensitivity of the method and pursues accuracy without considering the patient's clinical benefit. For example, when using biomarkers to predict the patient's disease, no matter which value is selected as the critical value, there will be the possibility of false positives and false negatives. Sometimes avoiding false positives benefits more, sometimes it is more hopeful of avoiding false negatives. Since both situations are unavoidable, then we want to find a way to maximize the net benefit. Therefore, to evaluate the clinical usefulness, whether our model can improve patient prognosis, decision curve analysis was conducted. DCA is a novel strategy to access the value of diagnostic tests that target various patient preferences to accept the risks of undertreatment and

overtreatment [32]. In our study, the decision curve indicated that the threshold probability is > 9%, the effect of using the selection model of random forest algorithm to predict LNM is greater than the scheme of "treatment none" or "treatment all."

First, one of the study limitations is genomics and radiomics have not been considered. Some studies have recently proved that genomics and radiomics can predict LNM in NSCLC [33–35], but we are unsure whether adding these factors will improve our model. Second, small sample size will lead to residual confounding or statistical fluctuation [36]. Another limitation is that the lack of external validation may limit the extrapolation of the model. Although all the results showed that the established model has satisfactory performance, it still needs to be externally validated in future research datasets from other centers.

Conclusion

In summary, this study developed a machine learning model that combines immunohistochemical characteristics and clinicopathological features, which may be useful to clinicians applied to facilitate the individualized prediction of preoperative LNM in NSCLC patients.

Abbreviations

LNM, lymph node metastasis; NSCLC, non-small cell lung cancer; AUROC, area under receiver operating characteristic curve; AUC, area under the receiver operating characteristic curve; CT, computed tomography; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; LASSO, least absolute shrinkage and selection operator algorithms; DCA, decision curve analysis; CK5/6, cytokeratin 5/6; CK7, cytokeratin 7; TTF-1,thyriod transcription factor-1; Napsin-A: aspartic proteinase napsin; P63, tumor protein p63.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

(I) Conception and design: C Mao, H Shen, M Cai, D Shen; (II) Administrative support: C Mao, H Shen; (III) Provision of study materials or patients: C Mao, H Shen; (IV) Collection and assembly of data: M Cai, D Shen; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Funding

None.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The study was approved by institutional ethics board of Nan fang hospital and individual consent for this retrospective analysis was waived.

Consent for publication

The consent to publish this manuscript has been obtained from all authors.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7–30.
2. Fossella F, Pereira JR, von Pawel J, et al. Randomized, Multinational, Phase III Study of Docetaxel Plus Platinum Combinations Versus Vinorelbine Plus Cisplatin for Advanced Non–Small-Cell Lung Cancer: The TAX 326 Study Group. J Clin Oncol. 2003;21(16):3016–24.
3. Fujii H, Horie S, Sukhbaatar A, et al. Treatment of false-negative metastatic lymph nodes by a lymphatic drug delivery system with 5-fluorouracil. Cancer Med-Us. 2019;8(5):2241–51.
4. Cheng Z, Yang P, Qu S, et al. Risk factors and management for early and late intrahepatic recurrence of solitary hepatocellular carcinoma after curative resection. Hpb. 2015;17(5):422–27.
5. Huang C, Yue J, Li Z, Li N, Zhao J, Qi D. Usefulness of the neutrophil-to-lymphocyte ratio in predicting lymph node metastasis in patients with non-small cell lung cancer. Tumor Biology. 2015;36(10):7581–89.
6. Barton JB, Langdale LA, Cummins JS, et al. The utility of routine preoperative computed tomography scanning in the management of veterans with colon cancer. The American Journal of Surgery. 2002;183(5):499–503.
7. Leufkens AM, van den Bosch MAAJ, van Leeuwen MS, Siersema PD. Diagnostic accuracy of computed tomography for colon cancer staging: A systematic review. *Scand J Gastroentero* 2011; 46(7–8): 887 – 94.
8. Flores RM, Akhurst T, Gonen M, Larson SM, Rusch VW. Positron emission tomography defines metastatic disease but not locoregional disease in patients with malignant pleural mesothelioma. The Journal of Thoracic Cardiovascular Surgery. 2003;126(1):11–5.
9. Erasmus JJ, Truong MT, Smythe WR, et al. Integrated computed tomography-positron emission tomography in patients with potentially resectable malignant pleural mesothelioma: Staging implications. The Journal of Thoracic Cardiovascular Surgery. 2005;129(6):1364–70.

10. Schouwink JH, Schultze Kool L, Rutgers EJ, et al. The value of chest computer tomography and cervical mediastinoscopy in the preoperative assessment of patients with malignant pleural mesothelioma. The Annals of Thoracic Surgery. 2003;75(6):1715–18.
11. Arita T, Matsumoto T, Kuramitsu T, et al. Is it possible to differentiate malignant mediastinal nodes from benign nodes by size? Reevaluation by CT, transesophageal echocardiography, and nodal specimen. Chest. 1996;110(4):1004–08.
12. Yasufuku K, Nakajima T, Motoori K, et al. Comparison of Endobronchial Ultrasound, Positron Emission Tomography, and CT for Lymph Node Staging of Lung Cancer. Chest. 2006;130(3):710–18.
13. Wang Memoli JS, El-Bayoumi E, Pastis NJ, et al. Using Endobronchial Ultrasound Features to Predict Lymph Node Metastasis in Patients With Lung Cancer. Chest. 2011;140(6):1550–56.
14. Kaseda K, Asakura K, Kazama A, Ozawa Y. Risk Factors for Predicting Occult Lymph Node Metastasis in Patients with Clinical Stage I Non-small Cell Lung Cancer Staged by Integrated Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography. World J Surg. 2016;40(12):2976–83.
15. Ding N, Mao Y, Gao S, et al. Predictors of lymph node metastasis and possible selective lymph node dissection in clinical stage IA non-small cell lung cancer. J Thorac Dis. 2018;10(7):4061–68.
16. Li X, Zhang H, Xing L, et al. Predictive Value of Primary Fluorine-18 Fluorodeoxyglucose Standard Uptake Value for a Better Choice of Systematic Nodal Dissection or Sampling in Clinical Stage IA Non–Small-Cell Lung Cancer. Clin Lung Cancer. 2013;14(5):568–73.
17. Ouyang M, Tang K, Xu M, Lin J, Li T, Zheng X. Prediction of Occult Lymph Node Metastasis Using Tumor-to-Blood Standardized Uptake Ratio and Metabolic Parameters in Clinical N0 Lung Adenocarcinoma. Clin Nucl Med. 2018;43(10):715–20.
18. Huang YQ, Liang CH, He L, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol. 2016;34(18):2157–64.
19. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). Circulation. 2015;131(2):211–19.
20. Tsutani Y, Murakami S, Miyata Y, Nakayama H, Yoshimura M, Okada M. Prediction of lymph node status in clinical stage IA squamous cell carcinoma of the lung. Eur J Cardio-Thorac. 2015;47(6):1022–26.
21. Song C, Kimura D, Sakai T, Tsushima T, Fukuda I. Novel approach for predicting occult lymph node metastasis in peripheral clinical stage I lung adenocarcinoma. J Thorac Dis. 2019;11(4):1410–20.
22. Hayashi T, Saito T, Fujimura T, et al. Galectin-4, a novel predictor for lymph node metastasis in lung adenocarcinoma. Plos One. 2013;8(12):e81883.
23. Yang X, Pan X, Liu H, et al. A new approach to predict lymph node metastasis in solid lung adenocarcinoma: a radiomics nomogram. J Thorac Dis. 2018;10(Suppl 7):807-19.
24. Huang YQ, Liang CH, He L, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol. 2016;34(18):2157–64.
25. Hsu C, Liu C, Tain Y, Kuo C, Lin Y. Machine Learning Model for Risk Prediction of Community-Acquired Acute Kidney Injury Hospitalization From Electronic Health Records: Development and Validation Study.

J Med Internet Res. 2020;22(8):e16903.
26. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics. 2012;28(10):1368–75.
27. Yang X, Pan X, Liu H, et al. A new approach to predict lymph node metastasis in solid lung adenocarcinoma: a radiomics nomogram. J Thorac Dis. 2018;10(Suppl 7):807-19.
28. Lu G, Chen L, Wu S, Feng Y, Lin T. Comprehensive Analysis of Tumor-Infiltrating Immune Cells and Relevant Therapeutic Strategy in Esophageal Cancer. *Dis Markers* 2020; 2020(1–12.
29. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *Bmc Med Inform Decis* 2020; 20(1).
30. Trainor P, DeFilippis A, Rai S. Evaluation of Classifier Performance for Multiclass Phenotype Discrimination in Untargeted Metabolomics. Metabolites. 2017;7(2):30.
31. Li Y, Li Z, Chen F, Liu Q, Peng Y, Chen M. A LASSO-derived risk model for long-term mortality in Chinese patients with acute coronary syndrome. *J Transl Med* 2020; 18(1).
32. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA. 2015;313(4):409–10.
33. Choi N, Son DS, Lee J, et al. The signature from messenger RNA expression profiling can predict lymph node metastasis with high accuracy for non-small cell lung cancer. J Thorac Oncol. 2006;1(7):622–28.
34. Yang X, Pan X, Liu H, et al. A new approach to predict lymph node metastasis in solid lung adenocarcinoma: a radiomics nomogram. J Thorac Dis. 2018;10(Suppl 7):807-19.
35. Cong M, Feng H, Ren JL, et al. Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer. Lung Cancer. 2020;139:73–9.
36. Song Y, Wang L, Pittas AG, et al. Blood 25-Hydroxy Vitamin D Levels and Incident Type 2 Diabetes: A meta-analysis of prospective studies. Diabetes Care. 2013;36(5):1422–28.

Figures

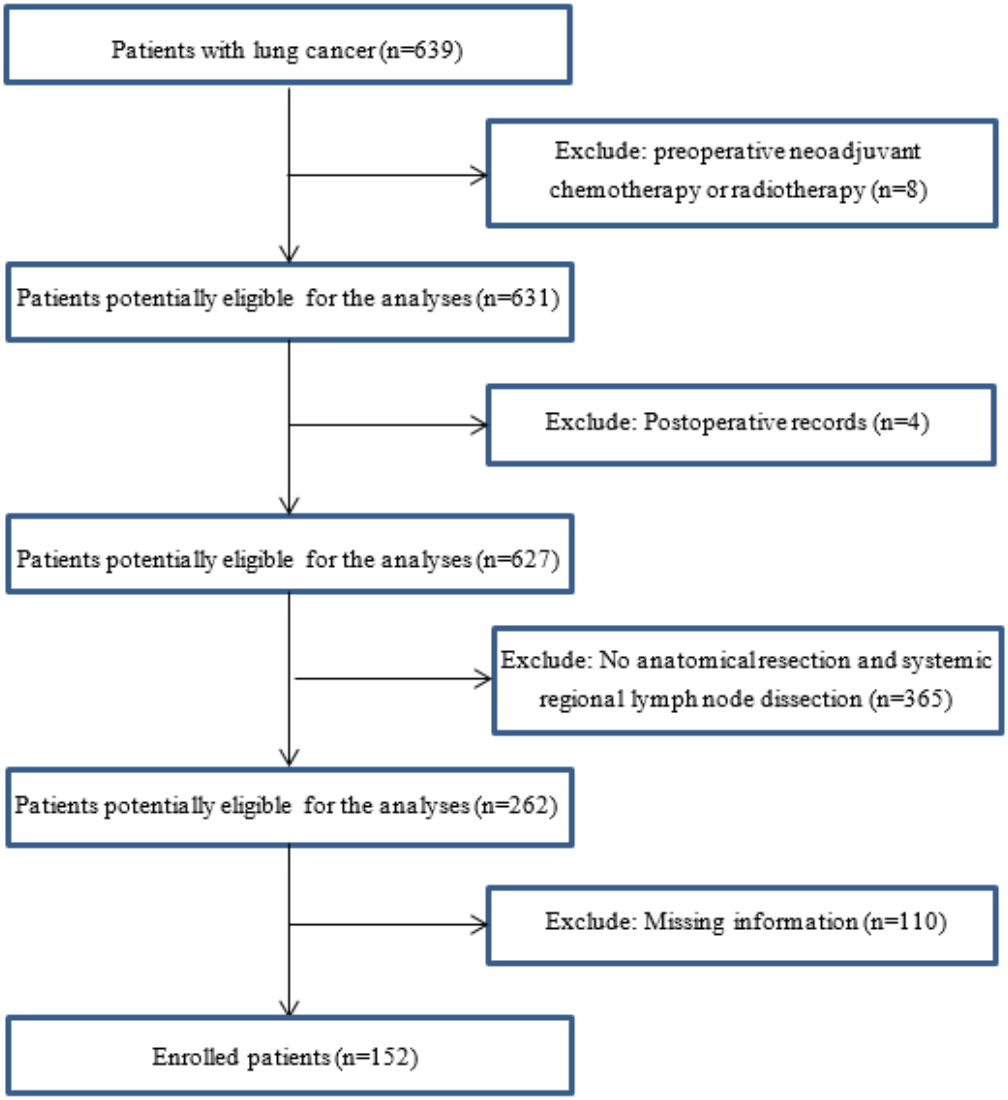

Figure 1

Flow diagram of selected patients.

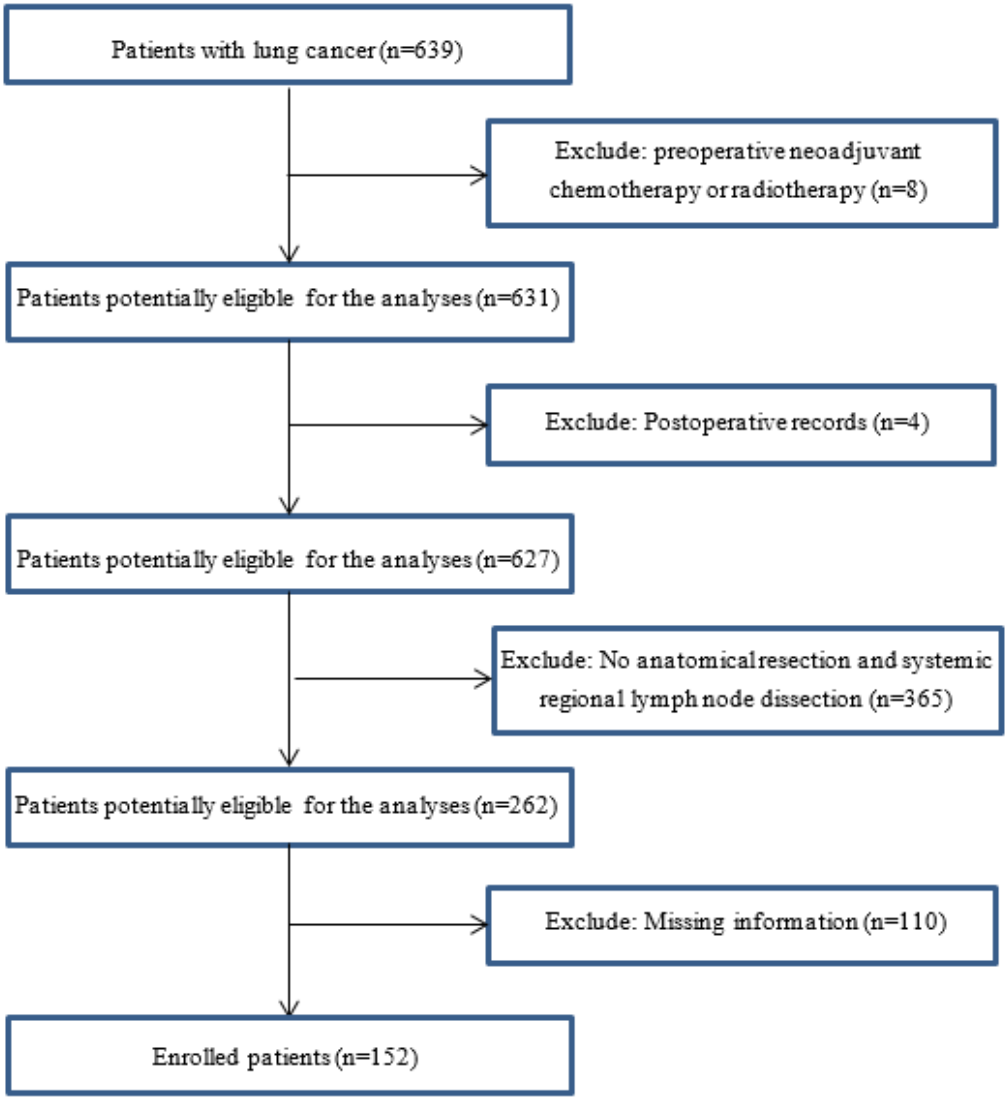

Figure 1

Flow diagram of selected patients.

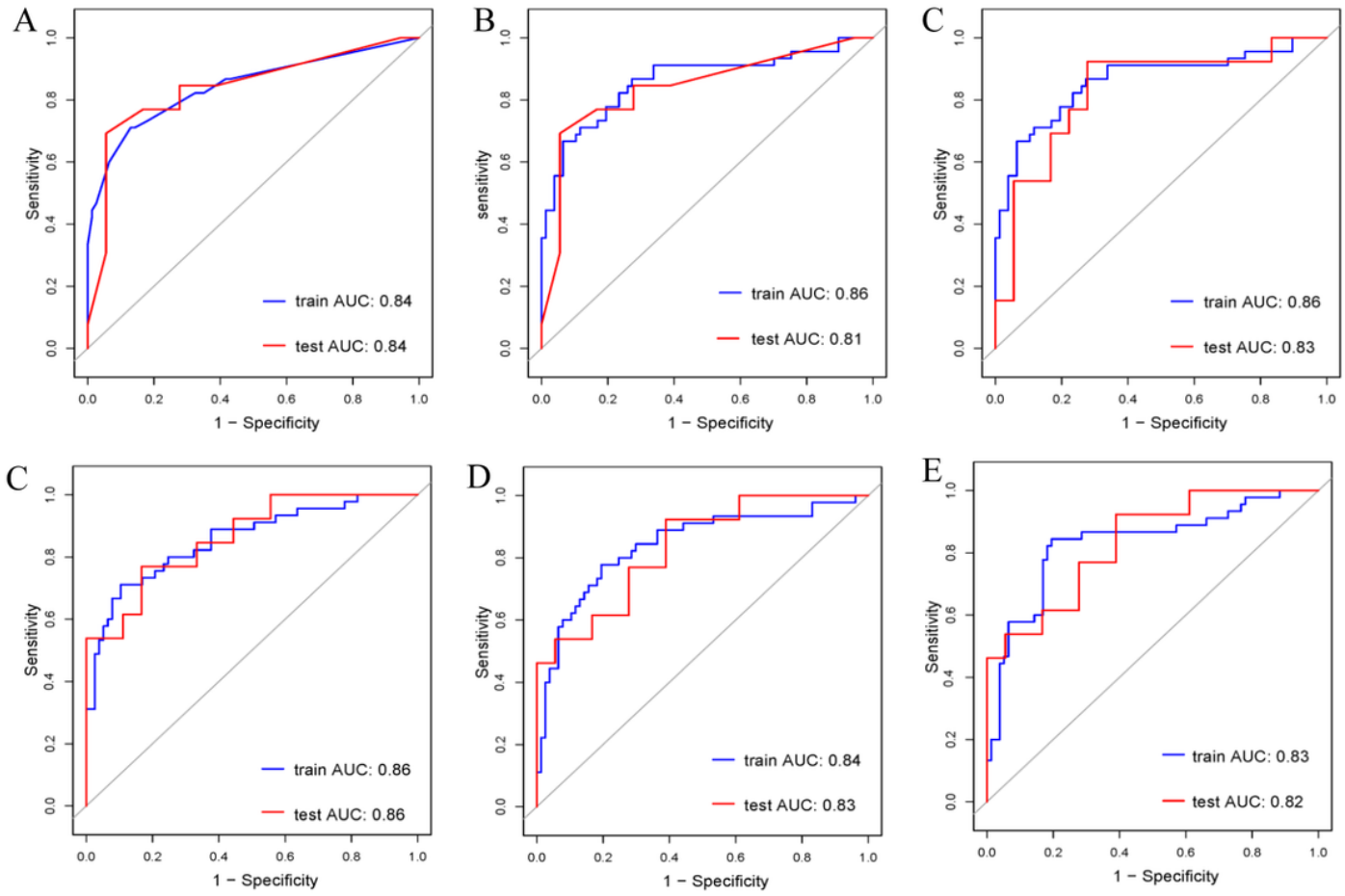

Figure 2

Area under receiver operating characteristic curves (AUROC) for each algorithm of training cohort and validation cohort. AUROC curves show the sensitivity on the y-axis and the 1-specificity on the X-axis. The blue line represents the AUC of the training set. The red line represents the AUC of the validation set. (A) AUROC curve of the random forest algorithm of the selection model in the training and validation cohort. (B) AUROC curve of the logistic regression algorithm of the selection model in the training and validation cohort. (C) AUROC curve of the support vector machine algorithm of the selection model in the training and validation cohort. (D) AUROC curve of the random forest algorithm of the full model in the training and validation cohort. (E) AUROC curve of the logistic regression algorithm of the full model in the training and validation cohort. (F) AUROC curve of the support vector machine algorithm of the full model in the training and validation cohort.

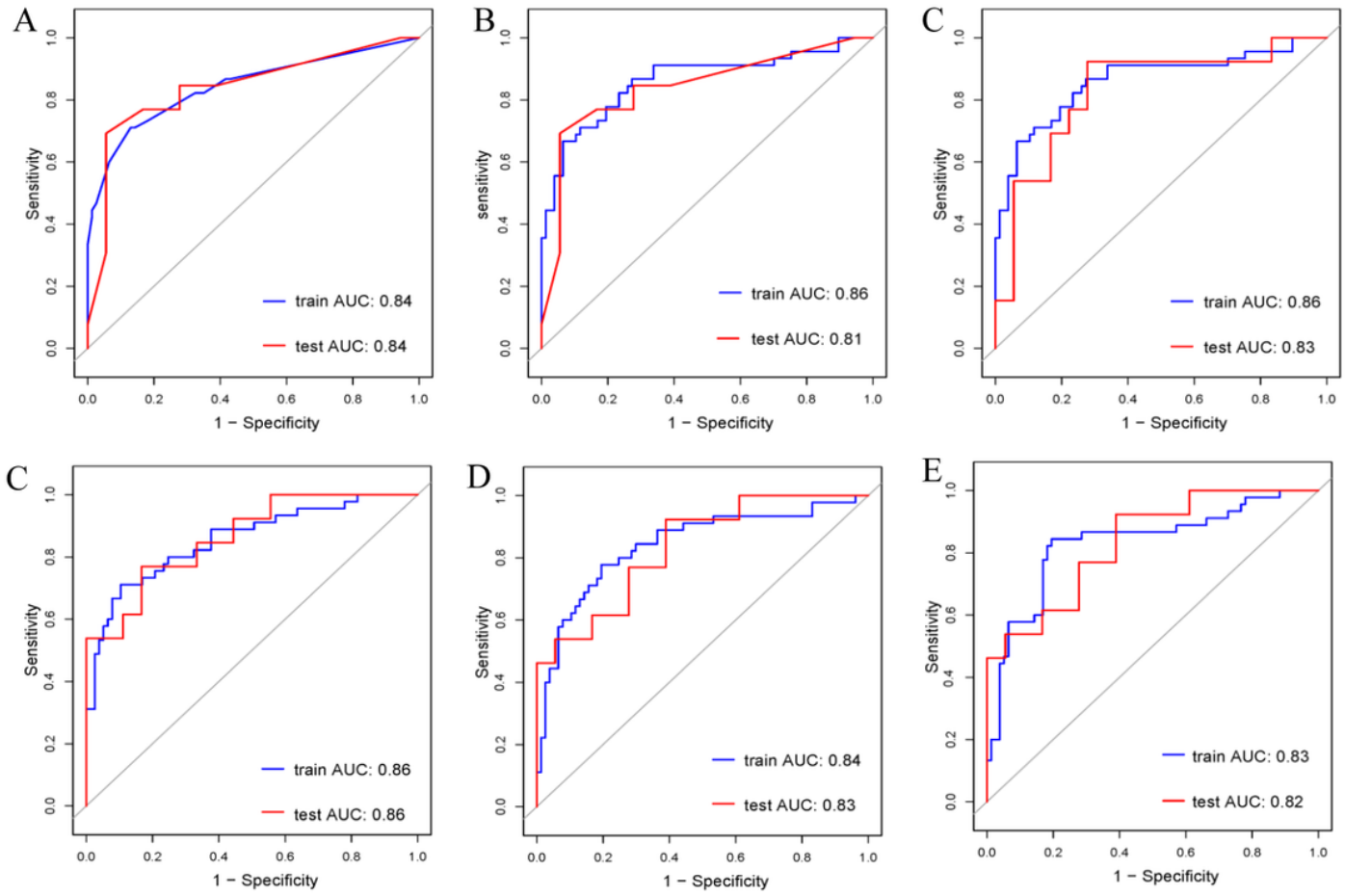

Figure 2

Area under receiver operating characteristic curves (AUROC) for each algorithm of training cohort and validation cohort. AUROC curves show the sensitivity on the y-axis and the 1-specificity on the X-axis. The blue line represents the AUC of the training set. The red line represents the AUC of the validation set. (A) AUROC curve of the random forest algorithm of the selection model in the training and validation cohort. (B) AUROC curve of the logistic regression algorithm of the selection model in the training and validation cohort. (C) AUROC curve of the support vector machine algorithm of the selection model in the training and validation cohort. (D) AUROC curve of the random forest algorithm of the full model in the training and validation cohort. (E) AUROC curve of the logistic regression algorithm of the full model in the training and validation cohort. (F) AUROC curve of the support vector machine algorithm of the full model in the training and validation cohort.

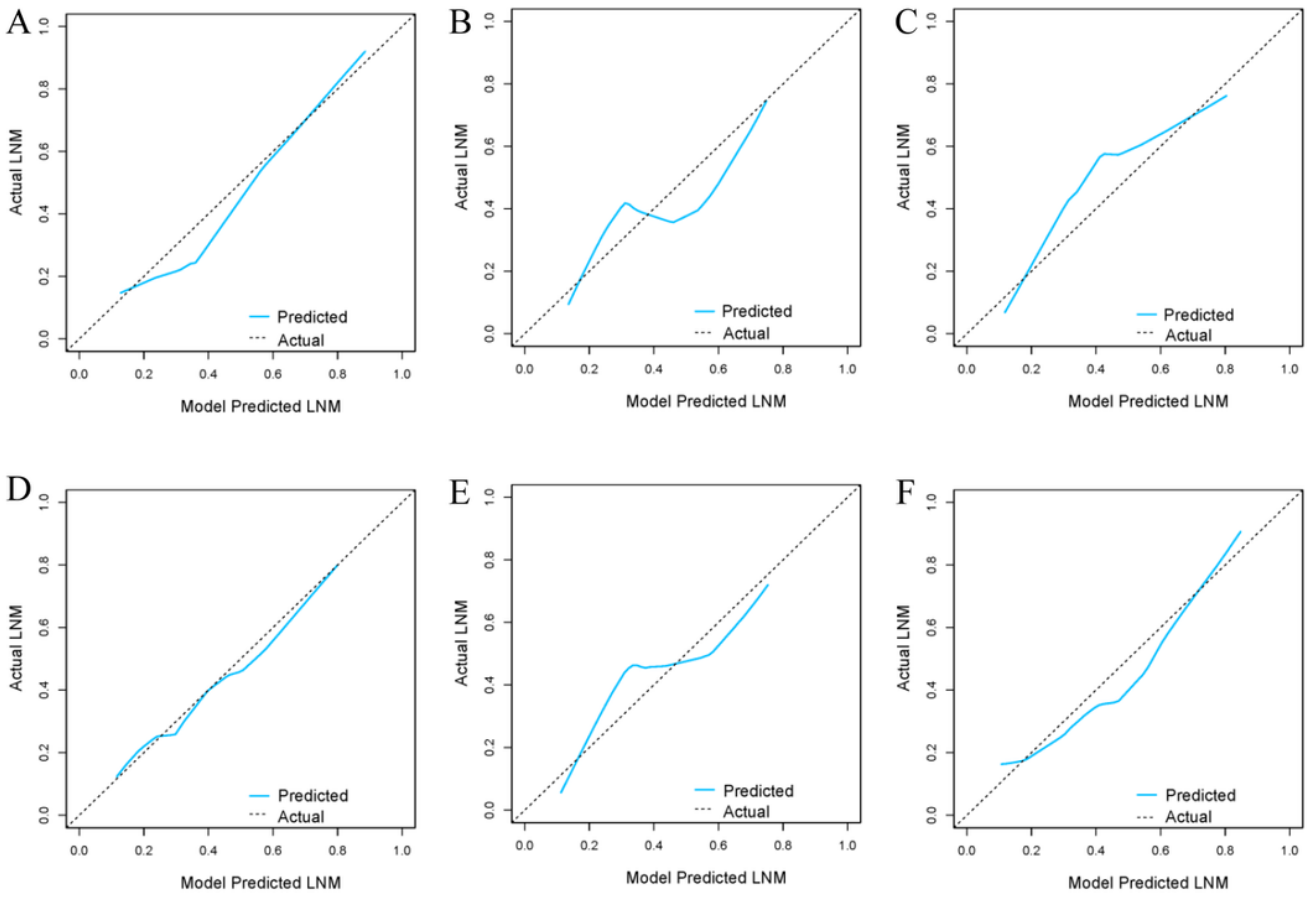

Figure 3

Calibration curve of models in the validation cohort. The y-axis shows the actual LNM probability. The x-axis shows the predicted LNM probability. The diagonal dashed line represents the ideal prediction of an ideal model. The blue solid line represents the performance of the model, and the closer the dotted line is to the diagonal point, the better the predictive performance of the model. (A) Calibration curve of the random forest algorithm of the selection model in the validation set. (B) Calibration curve of the logistic regression algorithm of the selection model in the validation set. (C) Calibration curve of the support vector machine algorithm of the selection model in the validation set. (D) Calibration curve of the random forest algorithm of the full model in the validation set. (E) Calibration curve of the logistic regression algorithm of the full model in the validation set. (F) Calibration curve of the support vector machine algorithm of the full model in the validation set.

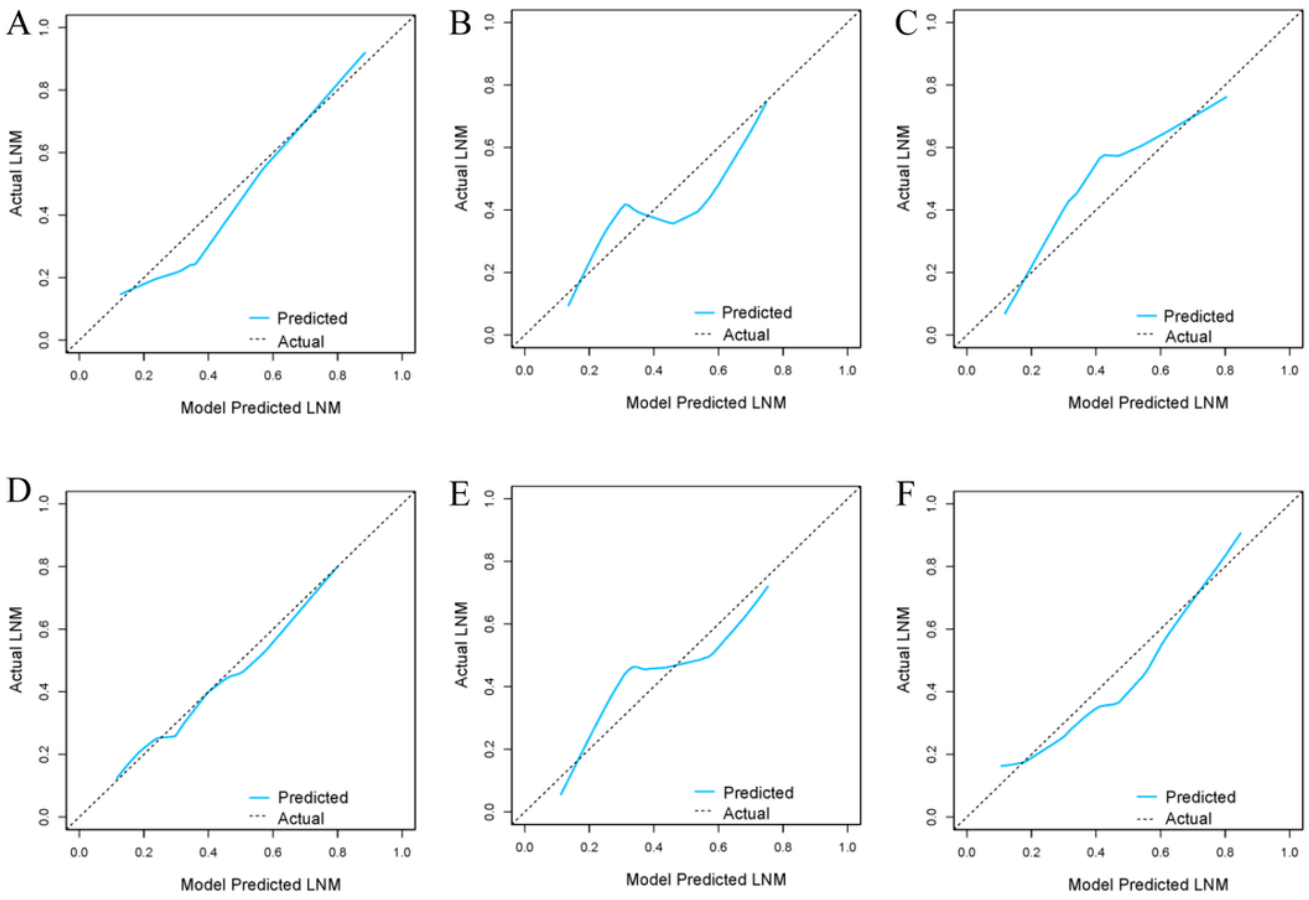

Figure 3

Calibration curve of models in the validation cohort. The y-axis shows the actual LNM probability. The x-axis shows the predicted LNM probability. The diagonal dashed line represents the ideal prediction of an ideal model. The blue solid line represents the performance of the model, and the closer the dotted line is to the diagonal point, the better the predictive performance of the model. (A) Calibration curve of the random forest algorithm of the selection model in the validation set. (B) Calibration curve of the logistic regression algorithm of the selection model in the validation set. (C) Calibration curve of the support vector machine algorithm of the selection model in the validation set. (D) Calibration curve of the random forest algorithm of the full model in the validation set. (E) Calibration curve of the logistic regression algorithm of the full model in the validation set. (F) Calibration curve of the support vector machine algorithm of the full model in the validation set.

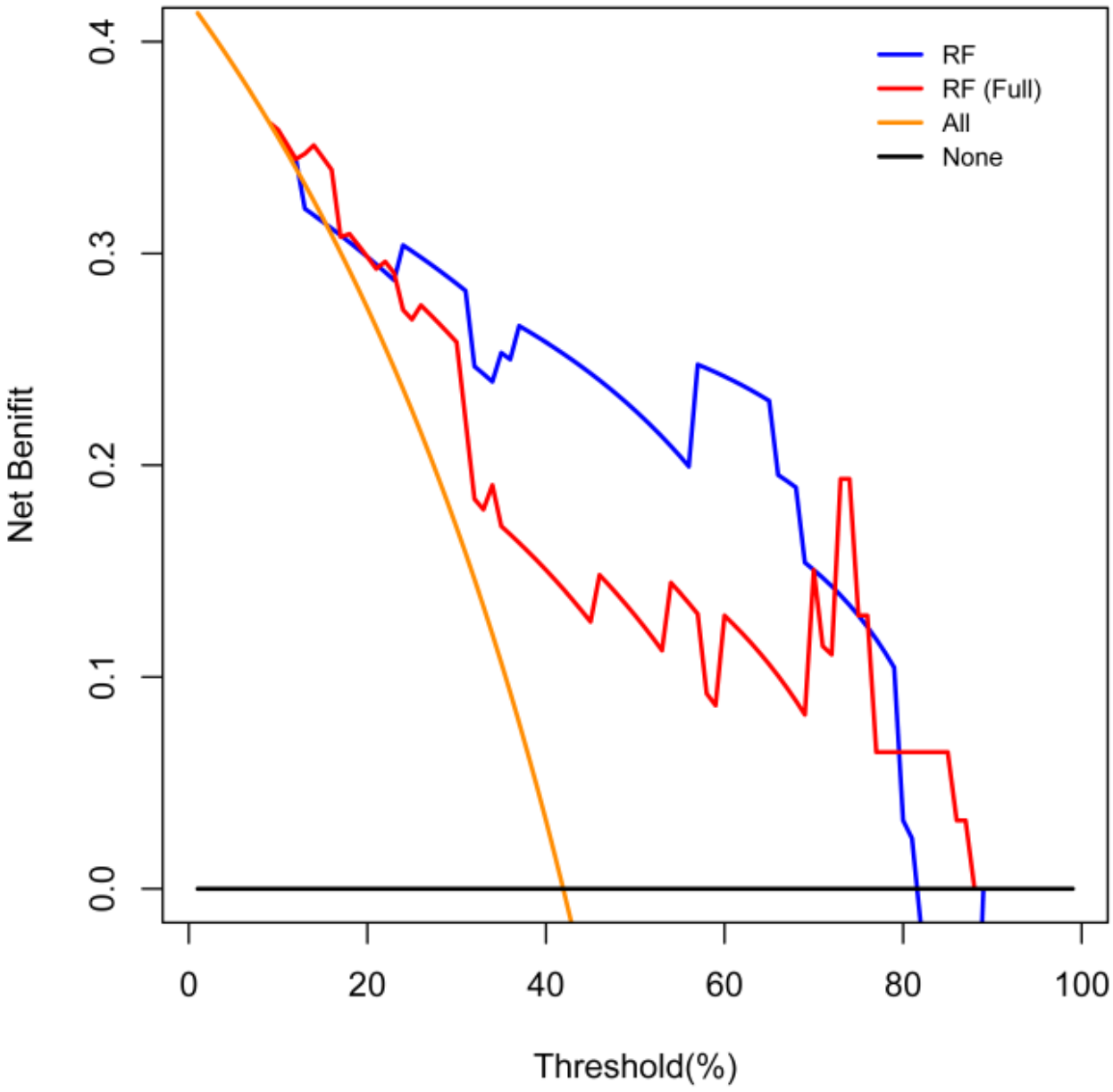

Figure 4

Decision curve analysis (DCA) for the random forest of the full model and the selection model. The Y-axis measures net benefit. The X-axis measures threshold probability. The dark origin line indicates the hypothesis that all patients accepted treatment. The black line indicates the hypothesis that no patients accepted treatment, so the net benefit of treatment must be 0. The blue line represents the random forest of the selection model. The red line represents the the random forest of the full model.

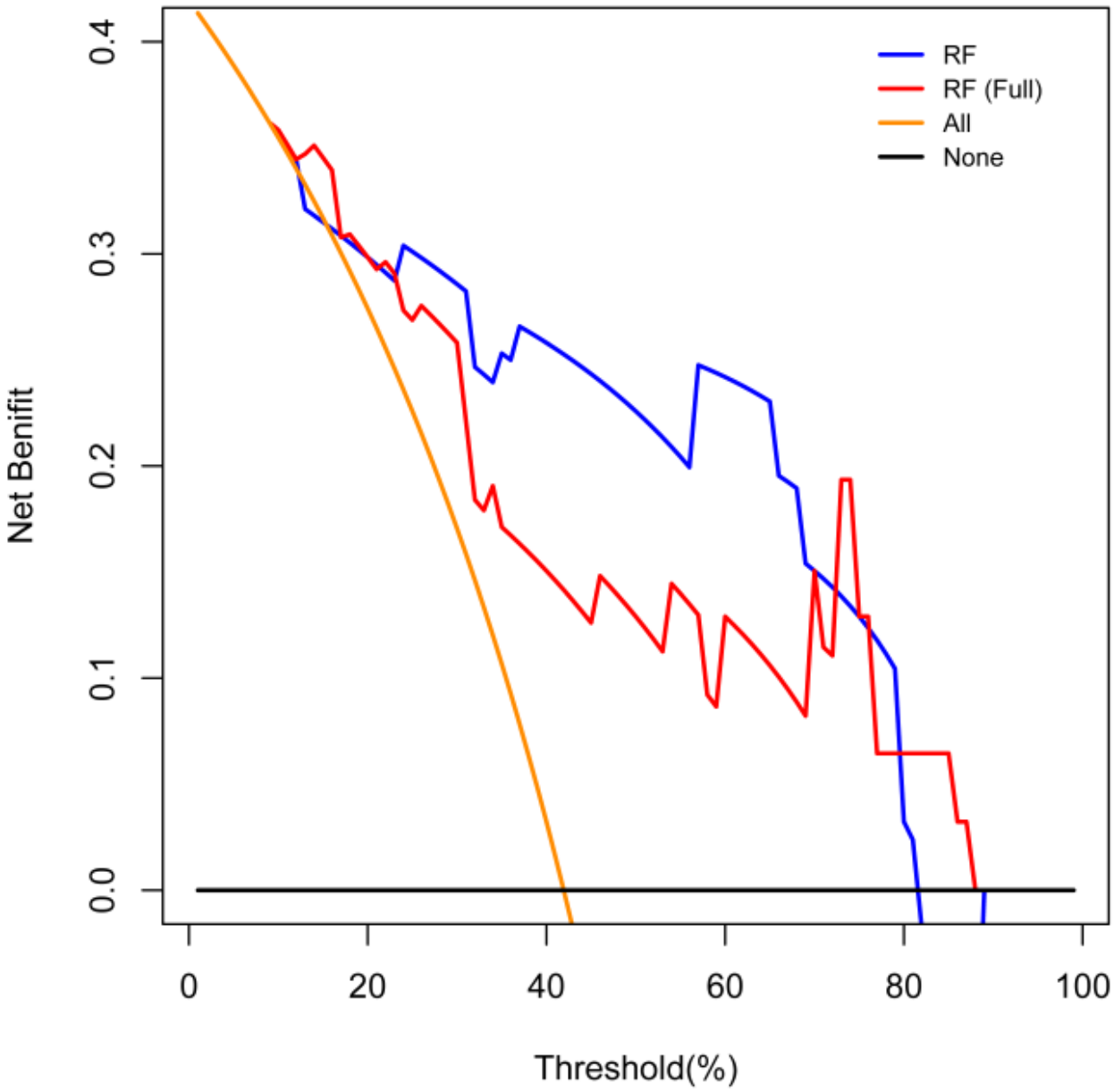

Figure 4

Decision curve analysis (DCA) for the random forest of the full model and the selection model. The Y-axis measures net benefit. The X-axis measures threshold probability. The dark origin line indicates the hypothesis that all patients accepted treatment. The black line indicates the hypothesis that no patients accepted treatment, so the net benefit of treatment must be 0. The blue line represents the random forest of the selection model. The red line represents the the random forest of the full model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryappendix.docx

- Supplementaryappendix.docx