Is it feasible to predict lymph node metastasis intraoperatively or postoperatively in early-stage lung adenocarcinoma: the application of machine learning algorithms?

Yijun Wu
Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, Ch

Yuming Chong
Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, China

Jianghao Liu
Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, China

Pancheng Wu
Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Yanyu Wang
Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Chang Han
Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, China

Liang Gong
Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, China

Xinyu Liu
Department of Radiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, China

Zhile Wang
Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. Peking Union Medical College, Eight-year MD program, Chinese Academy of Medical Sciences, Beijing, Ch

Naixin Liang
Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
Shanqing Li (lsq6768@163.com)
Peking Union Medical College Hospital

Research

Keywords: Non-small cell lung cancer, lymph node metastasis, predictive model, machine learning algorithm, decision curve analysis

Posted Date: May 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-29470/v1

Abstract

Background

Lymph node metastasis (LNM) status can be a critical decisive factor for clinical management of lung cancer. Accurately evaluating the risk of LNM during or after the surgery can be helpful for making clinical decisions. This study aims to incorporate clinicopathological characteristics to develop reliable machine learning (ML)-based models for predicting LNM in patients with early-stage lung adenocarcinoma.

Methods

A total of 709 lung adenocarcinoma patients with tumor size ≤ 2 cm were enrolled for analysis and modeling by multiple ML algorithms. The receiver operating characteristic (ROC) curve and decision curve were used for evaluating model's predictive performance and clinical usefulness. Feature selection based on potential models was performed to identify most-contributed predictive factors.

Results

LNM occurred in 11.3% (80/709) of patients with lung adenocarcinoma. Most models reached high areas under the ROC curve (AUCs) > 0.9. In the decision curve, all models performed better than the treat-all and treat-none lines. The random forest classifier (RFC) model, with a minimal number of 5 variables introduced (including carcinoembryonic antigen, solid component, micropapillary component, lymphovascular invasion and pleural invasion), was identified as the optimal model for predicting LNM, because of its excellent performance in both ROC and decision curves. The cost-efficient application of RFC model could precisely predict LNM during or after the operation of early-stage adenocarcinomas (sensitivity: 87.5%; specificity: 82.2%).

Conclusions

Incorporating clinicopathological characteristics, it is feasible to predict LNM intraoperatively or postoperatively by ML algorithms.

Trial registration: NA

Background

Lung cancer has been reported to be the most common cancer type worldwide and the leading cause of cancer death [1]. Among lung cancer cases that have various pathological characteristics, 80–85% of them can be categorized as non-small cell lung cancer (NSCLC) [2]. In the treatment of NSCLC, lymph node dissection (LND) during radical surgery is considered crucial [3]. A better understanding of lymph node metastasis (LNM) pattern aids to demarcate the extent of LND. Many studies focused on LNM in late-stage lung cancer, while LNM in small-size NSCLC should not be ignored as it could have an

incidence rate up to 10% [4, 5]. Moreover, occult LNM (OLNM) occurred not rarely in early-stage NSCLC [6–8], which might lead to a poor prognosis, especially for patients who received sublobar resection and sublevel excision of lymph nodes. Thus, it is more than necessary to precisely evaluate the risk of LNM intraoperatively and postoperatively, even in patients with no preoperatively suspected involvement of lymph nodes.

Machine learning (ML) generally defines an algorithm-based process that predicts outcome from large data files, presuming the existence of a pattern amidst the data that will identify the outcome. Comparing to traditional statistical models, ML predictive analysis has several benefits, including less outcomes required for each predictor, no requirement for specific hypothesis and allowance of interaction between variables [9, 10]. ML-based predictive analysis has been validly used in medical field [11, 12]. From the authors' perspective, there were very few studies that have reported the application of ML algorithms for evaluating the risk of LNM in lung cancer patients. This study aims to find validated ML models for the prediction of LNM in early-stage adenocarcinomas incorporating the clinical characteristics and postoperative histological patterns.

Subject And Methods

Study population

This study enrolled 709 NSCLC patients who has received lobectomy with systematic lymph node dissection at Peking Union Medical College Hospital (PUCMH) from January 2013 to December 2019. Enrolled patients had single foci NSCLC with maximum diameter ≤ 2 cm on CT. Patients who met any one of the following conditions were excluded: 1) diagnosed with small cell lung cancer; 2) diagnosed with multiple lung cancer; 3) preoperative radiotherapy or chemotherapy; 4) distant metastasis; 5) incomplete clinical information. The study was approved by the Institutional Review Board at PUCMH, Chinese Academy of Medical Science. All patients have signed written consent.

Clinicopathological characteristics

This study enrolled a total of 19 variables in three categories. Preoperative clinical characteristics included age, gender, smoking status, and serum carcinoembryonic antigen (CEA). Radiographical features were recorded from CT by one radiologist and two thoracic clinicians independently, which included tumor imaging density, tumor side, tumor maximum diameter and specific signs as spiculation, vessel convergence, lobulation and pleural indentation. Disagreement was solved by their consensus. Histologically, cancer lesions were divided into four subtypes, atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), microinvasive adenocarcinoma (MIA) and invasive adenocarcinoma (IA) [13]. For all tumor lesions, histological details were further examined by pathological experts at our hospital, which included the presence of papillary, micropapillary, solid, acinar and lepidic components. Additionally, lymphovascular invasion (LVI) and pleural invasion (PI) were also considered risk factors for LNM. Pathological staging was based on 8th edition TNM Classification for lung cancer [14].

Development and validation of ML-based models

Firstly, z-score normalization was preprocessed to code running for continuous variables except for multinomial Naïve Bayes (MNB) algorithm, to which min-max normalization was done [15]. For the prediction of lymph node metastasis, we applied two conventional models including logistic regression (LR) and MNB, and six representative supervised algorithms including adaptive boosting (ADB), artificial neural network (ANN), decision tree (DT), gradient boosting decision tree (GBDT), random forest classifier (RFC) and extreme gradient boosting (XGB) [16–21].

Overfitting, meaning model becomes too specific to be suitable for another dataset, is a common risk, especially when variable number is large. The cross-validation strategy have been proven effective for the avoidance of overfitting [22, 23]. In this study, enrolled patients were randomly and equally split into five datasets for 5-fold cross-validation. For each running action, one of datasets was used as the testing group and the remaining four as the training group. This process repeated 5 times for each algorithm to find the optimal models. The performance of ML-based models was evaluated by the area under the receiver operating characteristic (ROC) curve (AUC) for predictive ability and the decision curve for clinical usefulness.

Feature selection

A classifier-specific evaluator for feature contribution was applied to each model to select variables. The potential models with best predictive performance and clinical usefulness were picked up to identify predictive risk factors. A list of variables sequenced by predictive contribution to the models was returned. Lower rank indicated better relevance to the model.

Statistical analysis

Univariate analysis was performed using SPSS 25.0 (IBM, New York, USA). Normality for quantitative data was analyzed by Shapiro-Wilk test. Normal quantitative parameters were compared under Student's t test and written as mean ± standard deviation (SD), while non-normal quantitative parameters were compared under Mann-Whitney U test and written as median with interquartile (IQR). Pearson's Chi square test (or Fisher's exact test when necessary) was used to compare the distribution of categorical variables. ML-based models were developed using Python programming language (version 3.7). Decision curve analysis (DCA) was performed using R software (version 3.6.3). Statistical significance was considered as P value < 0.05 (two-side).

Results

Patient Characteristics

Table 1 lists the clinical characteristics of all 709 patients involved in this study. The patients aged from 51 to 64 with a median age of 58 years old. LNM was observed in 80 (11.3%) patients. The node-positive group had a median CEA concentration of 3.63 ng/ml, significantly higher than node-negative group,

indicating that a higher serum CEA level could be a risk factor of LNM. Additionally, a larger tumor size was significantly with LNM ($p < 0.001$). In terms of the radiologic characteristic of lung cancer foci, node-positive group and node-negative group were significantly different in tumor density ($p < 0.001$) and pleural indentation ($p = 0.02$), but not in spiculation ($p = 0.315$), vessel convergence ($p = 0.226$) or lobulation ($p = 0.154$). There was no pGGO cancer lesion in node-positive group. Further, analysis of clinicopathological features showed that the presence of micropapillary component ($p < 0.001$), solid component ($p < 0.001$), acinar component ($p = 0.001$), LVI ($p < 0.001$) and VPI ($p < 0.001$) could be possible risk factors of LNM, while the presence of lepidic component indicated LNM-free disease ($p < 0.001$). All node-positive patients were proved to be invasive adenocarcinomas by pathology.

Table 1
Univariate analysis predictors of lymph node metastasis

	Total	Lymph node status		p value
		pN_+	pN_0	
Patients	709	80 (11.3%)	629 (88.7%)	
Age (years)	58 [51–64]	58.5 [53–64]	58 [51–64]	0.744
Gender				
Male	256 (36.1%)	31 (38.8%)	225 (35.8%)	0.601
Female	453 (63.9%)	49 (61.3%)	404 (64.2%)	
Smoking history				
Yes	141 (19.9%)	19 (23.8%)	122 (17.2%)	0.358
No	568 (80.1%)	61 (76.3%)	507 (71.5%)	
Tumor side				
Right	447 (63.0%)	47 (58.8%)	400 (63.6%)	0.398
Left	262 (37.0%)	33 (41.3%)	229 (36.4%)	
CEA (ng/ml)	1.88 [1.21–2.85]	3.63 [1.97–6.82]	1.75 [1.17–2.56]	< 0.001
Tumor size (cm)	1.4 [1.0-1.7]	1.7 [1.5-2.0]	1.3 [1.0-1.6]	< 0.001
Tumor density				
Pure GGO	240 (33.9%)	0	240 (38.2%)	< 0.001
mGGO or solid	469 (66.1%)	80 (100%)	389 (61.8%)	
Spiculation				
Present	406 (57.3%)	50 (62.5%)	356 (56.6%)	0.315
Absent	303 (42.7%)	30 (37.5%)	273 (43.4%)	
Vessel convergence				
Present	162 (22.8%)	14 (17.5%)	148 (23.5%)	0.226
Absent	547 (77.2%)	66 (82.5%)	481 (76.5%)	
Lobulation				

pN_+: node-positive group; pN_0: node-negative group; mGGO: mixed ground glass opacity; AAH: atypical adenomatous hyperplasia; AIS: adenocarcinoma in situ; MIA: microinvasive adenocarcinoma; IA: invasive adenocarcinoma; LVI: lymphovascular invasion; PI: pleural invasion; CEA: carcinoembryonic antigen.

	Total	Lymph node status		p value
		pN_+	pN_0	
Present	276 (38.9%)	37 (46.3%)	239 (38.0%)	0.154
Absent	433 (61.1%)	43 (53.8%)	390 (62.0%)	
Pleural indentation				
Present	205 (28.9%)	32 (40.0%)	173 (27.5%)	0.020
Absent	504 (71.1%)	48 (60.0%)	456 (72.5%)	
Histology subtype				
AAH	8 (1.1%)	0	8 (1.3%)	< 0.001
AIS	45 (6.3%)	0	45 (7.2%)	
MIA	67 (9.4%)	0	67 (10.7%)	
IA	589 (83.1%)	80 (100%)	509 (80.9%)	
Papillary component				
Present	187 (26.4%)	25 (31.3%)	162 (25.8%)	0.293
Absent	522 (73.6%)	55 (68.8%)	467 (74.2%)	
Micropapillary component				
Present	119 (16.8%)	37 (46.3%)	82 (13.0%)	< 0.001
Absent	590 (83.2%)	43 (53.8%)	547 (87.0%)	
Solid component				
Present	63 (8.9%)	34 (42.5%)	29 (4.6%)	< 0.001
Absent	646 (91.1%)	46 (57.5%)	600 (95.4%)	
Acinar component				
Present	518 (73.1%)	71 (88.8%)	447 (71.1%)	0.001
Absent	191 (26.9%)	9 (11.3%)	182 (28.9%)	
Lepidic component				
Present	321 (45.3%)	8 (10.0%)	313 (49.8%)	< 0.001

pN_+: node-positive group; pN_0: node-negative group; mGGO: mixed ground glass opacity; AAH: atypical adenomatous hyperplasia; AIS: adenocarcinoma in situ; MIA: microinvasive adenocarcinoma; IA: invasive adenocarcinoma; LVI: lymphovascular invasion; PI: pleural invasion; CEA: carcinoembryonic antigen.

	Total	Lymph node status		p value
		pN_+	pN_0	
Absent	388 (54.7%)	72 (90.0%)	316 (50.2%)	
LVI				
Present	22 (3.1%)	12 (15.0%)	10 (1.6%)	< 0.001
Absent	687 (96.9%)	68 (85.0%)	619 (98.4%)	
PI				
Present	76 (10.7%)	25 (31.3%)	51 (8.1)	< 0.001
Absent	633 (89.3%)	55 (68.8%)	578 (91.9)	

pN_+: node-positive group; pN_0: node-negative group; mGGO: mixed ground glass opacity; AAH: atypical adenomatous hyperplasia; AIS: adenocarcinoma in situ; MIA: microinvasive adenocarcinoma; IA: invasive adenocarcinoma; LVI: lymphovascular invasion; PI: pleural invasion; CEA: carcinoembryonic antigen.

Predictive performance of ML-based models

Six supervised ML algorithms were used to develop efficient and reliable predictive models with 19 clinicopathological variables, and their predictive performance is illustrated in Fig. 1 and Table 2. Among them, RFC model gave the best predictive performance (AUC = 0.921, SD = 0.014), closely followed by GBDT (AUC = 0.919, SD = 0.014), XGBoost (AUC = 0.917, SD = 0.028) and ANN (AUC = 0.915, SD = 0.017). As for two conventional methods, LR also performed well (AUC = 0.935, SD = 0.013), while the performance of MNB (AUC = 0.876, SD = 0.023) was poor.

Table 2
Predictive performance of different models

Model	AUC			Number of optimal dimensions
	Mean	SD	95% CI	
ADB	0.895	0.017	0.861–0.929	10
ANN	0.915	0.014	0.887–0.942	7
DT	0.870	0.019	0.832–0.908	7
GBDT	0.919	0.014	0.891–0.947	14
LR	0.935	0.013	0.910–0.961	15
MNB	0.876	0.023	0.831–0.921	15
RFC	0.921	0.014	0.894–0.948	5
XGB	0.917	0.015	0.888–0.946	10

AUC: area under the receiver operating characteristic curve; ADB: adaptive boosting; ANN: artificial neural network; DT: decision tree; GBDT: gradient boosting decision tree; LR: logistic regression; MNB: multinomial naïve Bayes; RFC: random forest classifier; XGB: extreme gradient boosting; CEA: carcinoembryonic antigen

To further compare the clinical usefulness of models, DCA was performed (Fig. 2). Firstly, across almost the entire reasonable range of thresholds, all models performed better than the two extreme lines (treat-all and treat-none lines). Most of them showed similar net benefits under most circumstances except for DT model. At the thresholds < 0.28, LR presented slightly higher net benefits than other models. However, when the thresholds ≥ 0.28, RFC model performed best at most values of threshold probability. At the threshold range of 0-0.4, MNB performed almost worst among all models except DT. When thresholds > 0.4, the net benefits of ADB and ANN decreased sharply and were lower than other models except DT. Therefore, in addition to RFC and LR, XGB and GBDT, which showed stably higher net benefits than other four models, were also identified as potential models.

Variable importance

Based on four potential models (RFC, LR, XGB and GBDT) with great predictive performance and clinical usefulness, the top 10 important variables for LNM prediction and their rank are shown in Fig. 3. The solid component ranked top to be the most influential predictive factor, followed by CEA, pleural invasion, tumor imaging density, LVI, micropapillary component, histological type, acinar component, lepidic component and gender, respectively.

Development of a dynamic predictive application

RFC model was considered the optimal model because of its excellent performance in both ROC curve and decision curve, which reached a high AUC with the minimal number of variables introduced, including CEA, solid component, micropapillary component, LVI and PI. Thus, a dynamic application of RFC model with these 5 variables was developed for the convenience of clinicians and patients (https://nmgrmshinyappszzypumch.shinyapps.io/Pathology/) [24].

According to the application, the optimal cutoff point of risk probability to distinguish LNM (+) from LNM (-) was 13.85% (sensitivity: 87.5%; specificity: 82.2%). Figure 4 shows the risk probability distribution of all patients, which has been standardized by the following formula: (risk probability-13.85%)/standard deviation.

Discussion

LNM status is crucial for the treatment of early-stage NSCLC. To date, lobectomy plus systematic lymph node dissection is the standard management to achieve low recurrence rate and prolong survival [3, 25]. However, compared with selective LND or lymph node sampling, systematic LND could be more likely to cause a series of postoperative complications [26, 27]. On other occasions, sublobar resection including segmentectomy and wedge resection has been recommended for early-stage NSCLC patients, which showed similar survival outcome as lobectomy [28, 29] and could also preserve more lung function. However, the sublevel surgery as selective LND and sublobar resection could more possibly lead to tumor residual and thus a poor prognosis if LNM occurred. Moreover, occult LNM makes the situation more complicated. It has been estimated that the occurrence rate of OLNM could be between 10.8–17.2% among stage I lung cancer [30–32]. Patients with LNM might mistakenly undergo sublevel surgery, leading to a poor prognosis. For these patients, salvage management might be necessary. Therefore, more efforts should be given to accurately predict the LNM status during or after the operation.

Previous studies have revealed some possible predictive factors for LNM in NSCLC. Yu *et al* reported several independent risk factors including tumor size, pleural invasion, and carcinoembryonic antigen [33]. Pani *et al* found that histologic subtypes could be related to lymph node status [34]. Another similar study suggested different lymph node dissection strategy for different combination of various clinicopathological features and CEA concentration and albumin level [35]. These studies used uni- and multivariate analysis to reveal clinicopathological predictors for different LNM patterns. Our study, however, innovatively adopts ML algorithms to predict LNM by incorporating a large series of clinicopathological features. Among the predictive models, we found that RFC, GBDT, XGB, ANN all achieved AUC higher than 0.9, which was similar with LR model. However, in the decision curve, LR performed better than others at threshold < 0.28, while RFC performed the best at most points of thresholds ≥ 0.28 and always kept a stably high net benefit. It is noteworthy that all models performed significantly better than treat-all and treat-none lines, indicating our models had clinical practice values and patients could gain more benefits if corresponding managements were conducted according to the predictive outcome of these models.

Furthermore, based on four potential models we identified with great performance in both ROC and decision curves, the top ten variables were found out, including solid component, CEA, pleural invasion, tumor imaging density, LVI, micropapillary component, histological subtype, acinar component, lepidic component and gender. In addition to CEA and imaging density that have been reported by previous studies [4, 5], many histological features were also strongly related to the occurrence of LNM. Besides pleural invasion and LVI, histological details of growth such as the presence of solid, micropapillary and acinar components indicated high risk for LNM, while the presence of lepidic component could indicate LNM-free disease. In fact, these variables are conventionally not included in intraoperative pathology report. Our study emphasizes the importance of these histological features in the prediction of lymph node status. Thus, intraoperative pathology may be considered to include more detailed information about adenocarcinomas to further evaluate LNM risk, especially for patients who are hard to decide between lobectomy and sublobar resection. Importantly, the risk evaluation of LNM after surgery might be necessary for early-stage adenocarcinoma patients. For those who received sublobar resection or sublevel LND, the salvage management and close follow-up could be required if a high risk for LNM was observed based on our ML model.

In recent years, predicting metastasis with machine learning algorithms, as a promising alternative for other invasive or noninvasive diagnostic method, has been proven to be feasible in lung adenocarcinoma and colorectal cancer [11, 12]. These studies predicted on CT image and histologic evidence and obtained satisfying results. However, considered the sample size in the two study was not large, the validity of machine learning prediction needs to be further confirmed on a larger NSCLC patient population. Another methodological problem remained to be further explained is that the false-positive and false-negative rate need to be low enough to achieve good clinical utility. High AUC in ROC represents high predictive accuracy but does not necessary prove good clinical utility, because false-positive or false-negative results could reduce net benefit [36]. To seek for a model that has high predictive accuracy and net benefit, we adopted DCA which has been widely proven to be efficiently and interpretable in the evaluation of clinical utility [37]. From the decision curve, it was clear that RFC has the highest net benefit across the longest stable range of clinically reasonable preferences.

To further enhance the clinical usefulness of our study, a dynamic application of RFC model with 5 clinicopathological variables introduced was developed. So, clinicians and patients worldwide can benefit from our study and evaluate the LNM risk easily. The node-positive patients could be precisely identified by the RFC application (sensitivity: 87.5%; specificity: 82.2%; Fig. 4).

This study is not without limitation. The nature of retrospective analysis inevitably causes data acquisition bias. Additionally, the enrolled patients are from a single center and share an ethnicity. Future study is expected to validate the predictive performance of RFC model and more possible clinicopathological variables in a multicenter population.

Conclusions

This study comprehensively evaluated various ML-based predictive models and identified RFC model as the optimal one that accurately predicted LNM in early-stage adenocarcinomas. By feature selection, some clinicopathological characteristics were found to be strongly related to LNM. The pGGO or non-invasive adenocarcinoma (AAH, AIS and MIA) cancer lesion might indicate LNM-free disease, which also consisted with the presence of lepidic component. The application of RFC model was developed with great predictive ability and clinical usefulness. Thus, it can be feasible to evaluate the risk of LNM in patients with early-stage adenocarcinoma during or after the operation for clinical decision-making.

Abbreviations

NSCLC: non-small cell lung cancer; LND: lymph node dissection (LND); LNM: lymph node metastasis; ML: Machine learning: PUMCH: Peking Union Medical College Hospital; CEA: carcinoembryonic antigen; AAH: atypical adenomatous hyperplasia; AIS: adenocarcinoma in situ; MIA: microinvasive adenocarcinoma; IA: invasive adenocarcinoma; LVI: lymphovascular invasion; PI: pleural invasion; MNB: multinomial Naïve Bayes; LR: logistic regression; ADB: adaptive boosting; ANN: artificial neural network; DT: decision tree; GBDT: gradient boosting decision tree; RFC: random forest classifier; XGB: extreme gradient boosting; ROC: receiver operating characteristic; AUC: area under curve; SD: standard deviation; IQR: interquartile; DCA: Decision curve analysis; pN+: node-positive group; pN0: node-negative group; mGGO: mixed ground glass opacity.

Declarations

Ethics approval and consent to participate

The study was approved by the Institutional Review Board at PUCMH, Chinese Academy of Medical Science.

Consent for publication

Informed consent in written form has been received from all patients.

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by CAMS Innovation Fund for Medical Sciences (CIFMS), (2017-12M-1-009; 2019-I2M-1-001).

Authors' contributions

YJW and JHL analyzed and interpreted the data; YJW and YMC wrote the manuscript; YMC and PCW performed the statistical analysis; YYW, CH, LG, XYL and ZLW collected the data; NXL and SQL supervised the study. All authors read and approved the final manuscript.

Acknowledgements

We would like to give our sincere thanks to Professor Hongsheng Liu, Yushang Cui, Zhijun Han and Zhili Cao for their contributions to the clinical work.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018, 68:394-424.
2. Lonardo F, Li X, Kaplun A, Soubani A, Sethi S, Gadgeel S, Sheng S: The natural tumor suppressor protein maspin and potential application in non small cell lung cancer. *Curr Pharm Des* 2010, 16:1877-1881.
3. De Leyn P, Dooms C, Kuzdzal J, Lardinois D, Passlick B, Rami-Porta R, Turna A, Van Schil P, Venuta F, Waller D, et al: Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer. *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* 2014, 45:787-798.
4. E P, G K, X Z, B U, D J, C G, T P, J K, thoracic SSJTJo, surgery c: Factors associated with nodal metastasis in 2-centimeter or less non-small cell lung cancer. 2020, 159:1088-1096.e1081.
5. X Y, Y L, C S, cancer HBJT: Risk factors of lymph node metastasis in patients with non-small cell lung cancer ≤ 2 cm in size: A monocentric population-based analysis. 2018, 9:3-9.
6. K K, K A, A K, surgery OYJWjo: 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. 2016, 40:2976-2983.
7. SY P, JK Y, KJ P, Society LSJCitopotICI: Prediction of occult lymph node metastasis using volume-based PET parameters in small-sized peripheral non-small cell lung cancer. 2015, 15:21.
8. Y M, SY C, JK P, surgery LKJWjo: Risk Factors for Occult Lymph Node Metastasis in Peripheral Non-Small Cell Lung Cancer with Invasive Component Size 3 cm or Less. 2020.
9. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ: Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med* 2016, 35:1159-1177.

10. Waljee AK, Higgins PD: Machine learning in medicine: a primer for physicians. *Am J Gastroenterol* 2010, 105:1224-1226.
11. Zhong Y, Yuan M, Zhang T, Zhang YD, Li H, Yu TF: Radiomics Approach to Prediction of Occult Mediastinal Lymph Node Metastasis of Lung Adenocarcinoma. *AJR Am J Roentgenol* 2018, 211:109-113.
12. Takamatsu M, Yamamoto N, Kawachi H, Chino A, Saito S, Ueno M, Ishikawa Y, Takazawa Y, Takeuchi K: Prediction of early colorectal cancer metastasis by machine learning using digital slide images. *Comput Methods Programs Biomed* 2019, 178:155-161.
13. WD T, E B, M N, AG N, KR G, Y Y, DG B, CA P, GJ R, PE VS, et al: International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. 2011, 6:244-285.
14. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, Nicholson AG, Groome P, Mitchell A, Bolejack V: The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* 2016, 11:39-51.
15. Shalabi. LA, Shaaban. Z, Kasasbeh. B: Data Mining: A Preprocessing Engine. *J Comput Sci* 2006, 2:735-739.
16. Ngiam KY, Khor IW: Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019, 20:e262-e273.
17. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS: Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in Bioinformatics* 2016, 17:33-42.
18. L. B: Random forests. *Mach Learn* 2001, 45:5-32.
19. Y. F, R. S, N. A: A short introduction to boosting. *Jinko Chino Gakkaishi* 1999, 14:1612.
20. Y. F, L. M: The alternating decision tree learning algorithm. *ICML* 1999, 99:124-133.
21. T. C, C. G: Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM* 2016:pp 785-794.
22. Cook JA, Ranstam J: Overfitting. *Br J Surg* 2016, 103:1814.
23. Jung Y: Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics* 2018, 30:197-215.
24. Calculation tool for predicting the risk of lymph node metastasis in lung adenocarcinoma. https://nmgrmshinyappszzypumch.shinyapps.io/Pathology/. Accessed May 15 2020.
25. Ginsberg RJ, Rubinstein LV: Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. Lung Cancer Study Group. *The Annals of thoracic surgery* 1995, 60.
26. Han H, Zhao Y, Chen H: Selective versus systematic lymph node dissection (other than sampling) for clinical N2-negative non-small cell lung cancer: a meta-analysis of observational studies. *J Thorac Dis* 2018, 10:3428-3435.

27. Okada M, Sakamoto T, Yuki T, Mimura T, Miyoshi K, Tsubota N: Selective mediastinal lymphadenectomy for clinico-surgical stage I non-small cell lung cancer. *Ann Thorac Surg* 2006, 81:1028-1032.
28. Cao J, Yuan P, Wang Y, Xu J, Yuan X, Wang Z, Lv W, Hu J: Survival Rates After Lobectomy, Segmentectomy, and Wedge Resection for Non-Small Cell Lung Cancer. *Ann Thorac Surg* 2018, 105:1483-1491.
29. Altorki NK, Yip R, Hanaoka T, Bauer T, Aye R, Kohman L, Sheppard B, Thurer R, Andaz S, Smith M, et al: Sublobar resection is equivalent to lobectomy for clinical stage 1A lung cancer in solid nodules. *J Thorac Cardiovasc Surg* 2014, 147:754-762; Discussion 762-754.
30. Park SY, Yoon JK, Park KJ, Lee SJ: Prediction of occult lymph node metastasis using volume-based PET parameters in small-sized peripheral non-small cell lung cancer. *Cancer Imaging* 2015, 15:21.
31. 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, 40:2976-2983.
32. Moon Y, Choi SY, Park JK, Lee KY: Risk Factors for Occult Lymph Node Metastasis in Peripheral Non-Small Cell Lung Cancer with Invasive Component Size 3 cm or Less. *World J Surg* 2020, 44:1658-1665.
33. Yu X, Li Y, Shi C, Han B: Risk factors of lymph node metastasis in patients with non-small cell lung cancer </= 2 cm in size: A monocentric population-based analysis. *Thorac Cancer* 2018, 9:3-9.
34. Pani E, Kennedy G, Zheng X, Ukert B, Jarrar D, Gaughan C, Pechet T, Kucharczuk J, Singhal S: Factors associated with nodal metastasis in 2-centimeter or less non-small cell lung cancer. *J Thorac Cardiovasc Surg* 2020, 159:1088-1096 e1081.
35. Zhao F, Zhen FX, Zhou Y, Huang CJ, Yu Y, Li J, Li QF, Zhu CX, Yang XY, You SH, et al: Clinicopathologic predictors of metastasis of different regional lymph nodes in patients intraoperatively diagnosed with stage-I non-small cell lung cancer. *BMC Cancer* 2019, 19:444.
36. Zhang Z, Rousson V, Lee WC, Ferdynus C, Chen M, Qian X, Guo Y, written on behalf of AMEB-DCTCG: Decision curve analysis: a technical note. *Ann Transl Med* 2018, 6:308.
37. Vickers AJ, van Calster B, Steyerberg EW: A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019, 3:18.

Figures

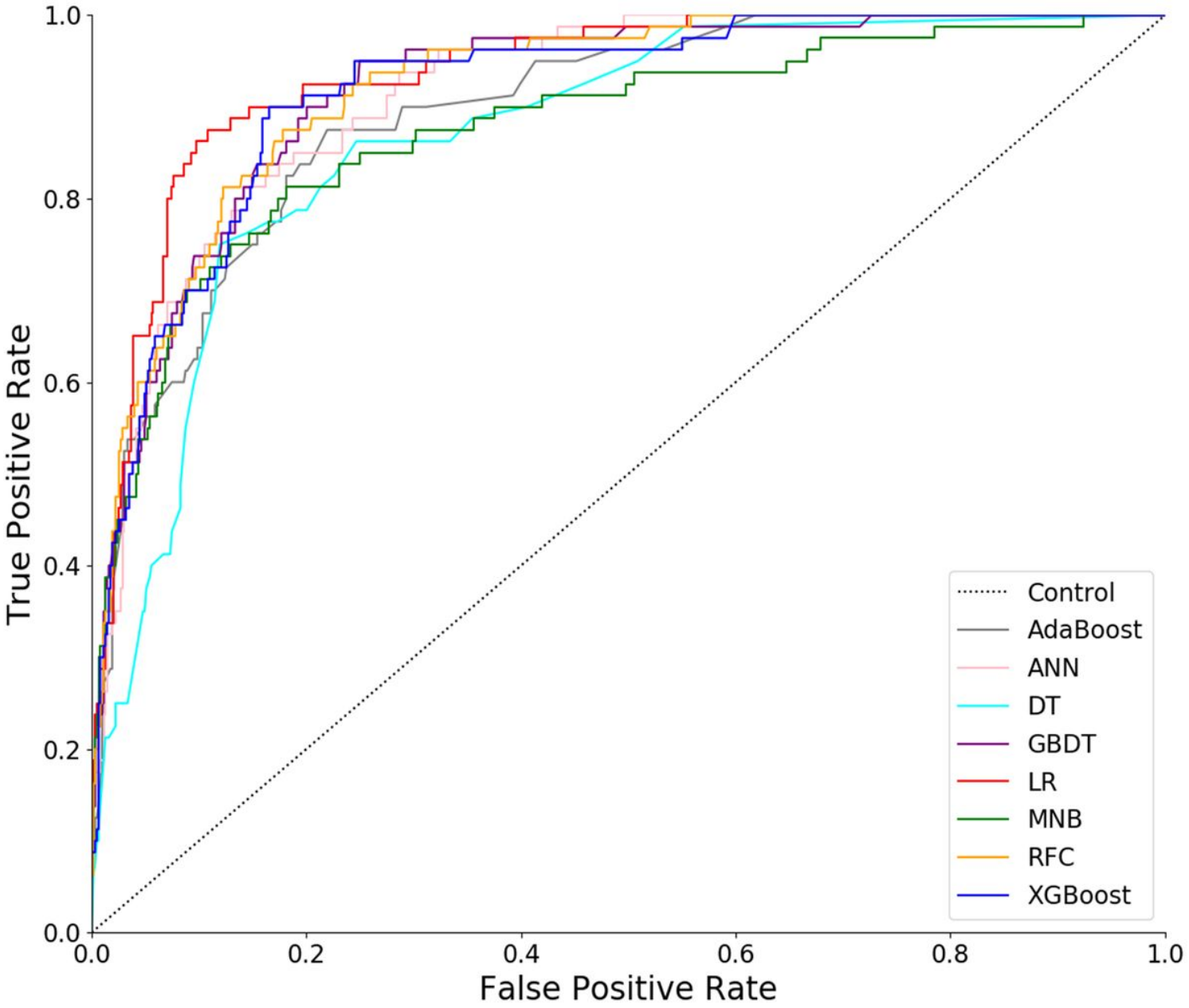

Figure 1

Receiver operating characteristic (ROC) curve for different predictive models. AdaBoost: adaptive boosting; ANN: artificial neural network; DT: decision tree; GBDT: gradient boosting decision tree; LR: logistic regression; MNB: multinomial naïve Bayes; RFC: random forest classifier; XGBoost: extreme gradient boosting.

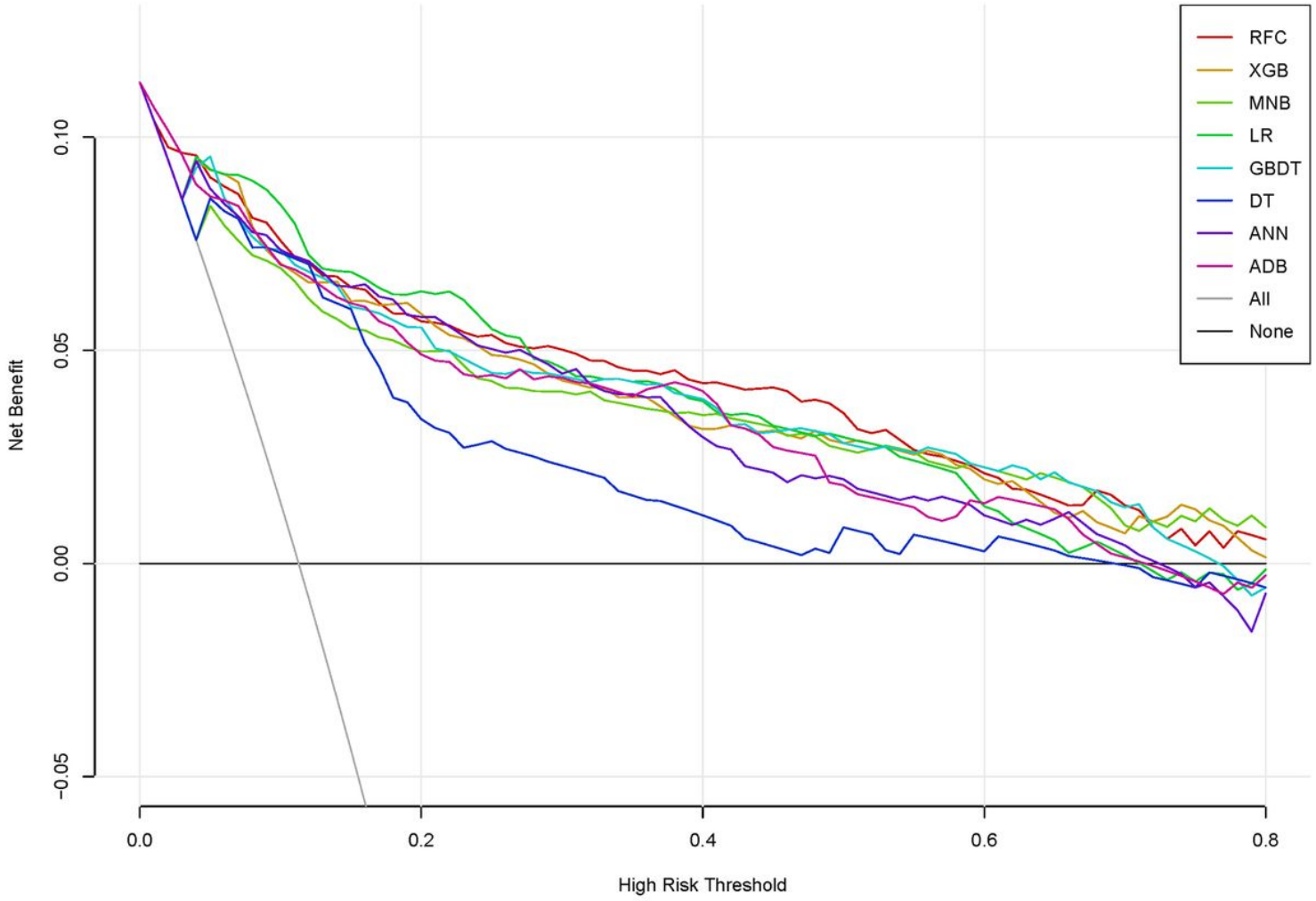

Figure 2

Decision curve for predictive models. AdaBoost: adaptive boosting; ANN: artificial neural network; DT: decision tree; GBDT: gradient boosting decision tree; LR: logistic regression; MNB: multinomial naïve Bayes; RFC: random forest classifier; XGBoost: extreme gradient boosting.

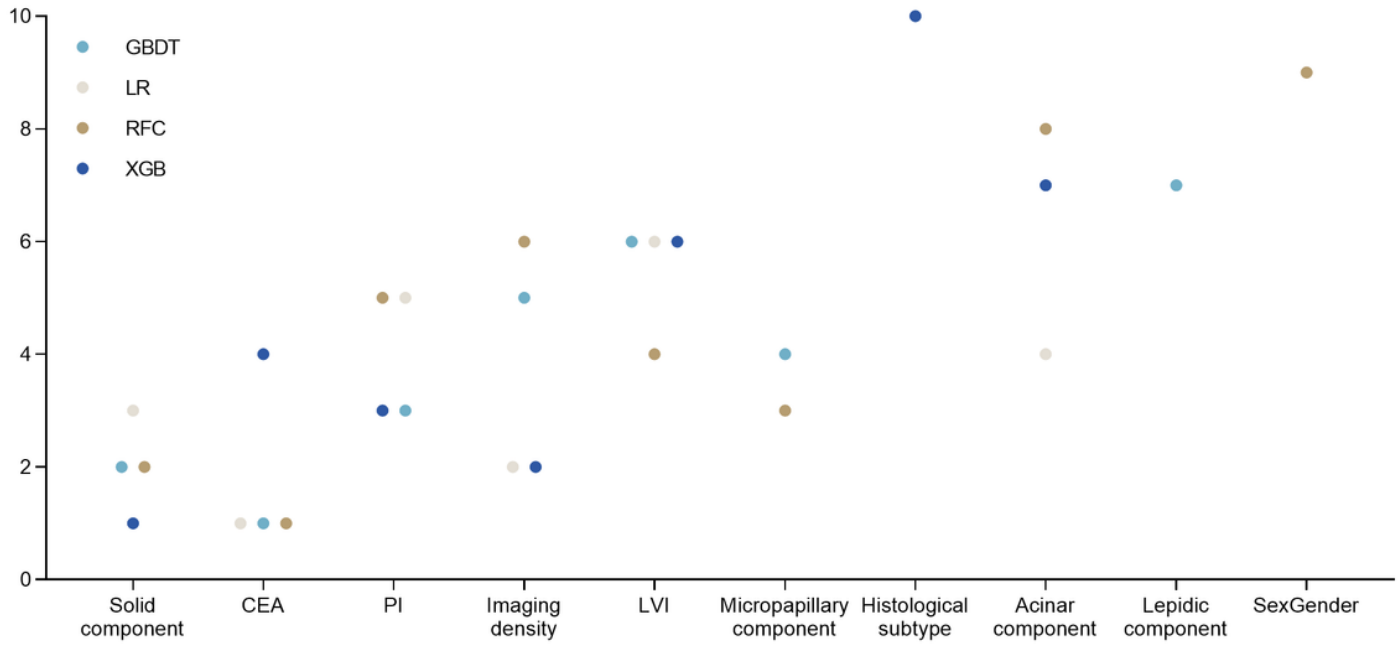

Figure 3

Top 10 important variables for predicting lymph node metastasis. LVI: lymphovascular invasion; PI: pleural invasion; CEA: carcinoembryonic antigen.

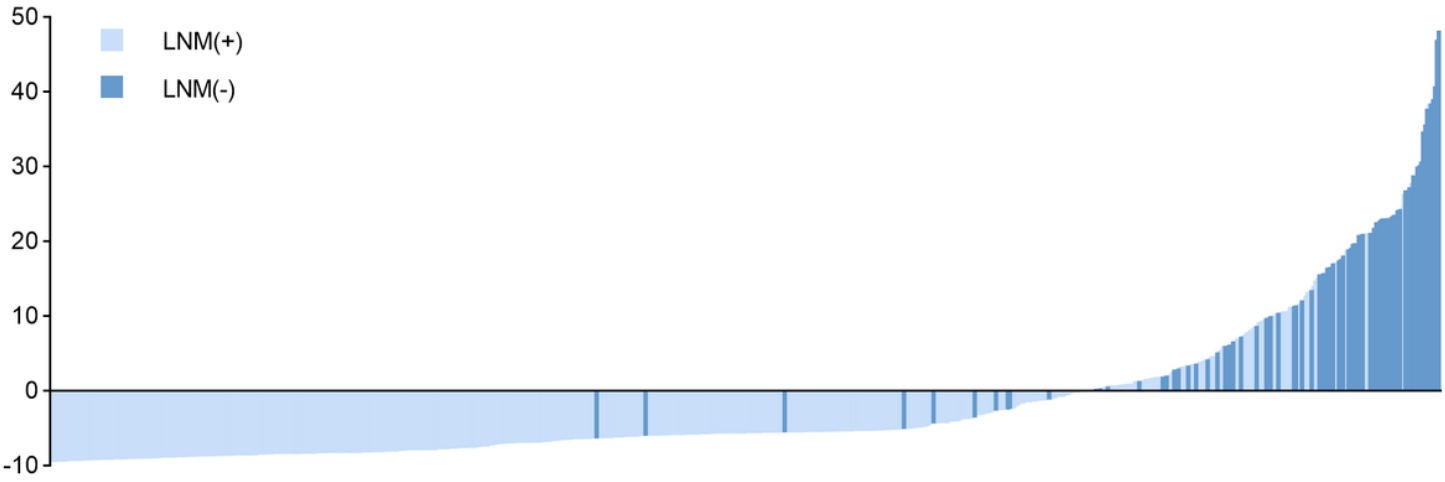

Figure 4

The standardized risk probability of each patient based on the random forest classifier (RFC) model. X-axis: each patient; Y-axis: the standardized risk probability.