IOP Publishing Phys. Med. Biol. 65 (2020) 055012 (13pp) https://doi.org/10.1088/1361-6560/ab6f98

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

A novel radiomic nomogram for predicting epidermal growth factor receptor mutation in peripheral lung adenocarcinoma

Xiaoqian Lu[1,5], Mingyang Li[2,5], Huimao Zhang[1,5], Shucheng Hua[3], Fanyang Meng[1], Hualin Yang[2], Xueyan Li[2,4] and Dianbo Cao[1,4]

[1] Department of Radiology, the First Hospital of Jilin University, 130021 Changchun, People's Republic of China
[2] State Key Laboratory of Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, 130012 Changchun, People's Republic of China
[3] Department of Pneumology, the First Hospital of Jilin University, 130021 Changchun, People's Republic of China
[4] Author to whom any correspondence should be addressed.
[5] These authors contributed equally to this work.

E-mail: 1196133391@qq.com (X Lu), limyctw@gmail.com (M Li), huimaozhanglinda@163.com (H Zhang), huasch@126.com (S Hua), mengjdyy1802@163.com (F Meng), 1739027586@qq.com (H Yang), leexy@jlu.edu.cn (X Li) and caotianbo@126.com (D Cao)

RECEIVED
13 August 2019

REVISED
18 January 2020

ACCEPTED FOR PUBLICATION
24 January 2020

PUBLISHED
6 March 2020

Keywords: epidermal growth factor receptor (EGFR), lung adenocarcinoma, computed tomography (CT), radiomics

Abstract

To predict the epidermal growth factor receptor (EGFR) mutation status in patients with lung adenocarcinoma using quantitative radiomic biomarkers and semantic features.

We analyzed the computed tomography (CT) images and medical record data of 104 patients with lung adenocarcinoma who underwent surgical excision and EGFR mutation detection from 2016 to 2018 at our center. CT radiomic and semantic features that reflect the tumors' heterogeneity and phenotype were extracted from preoperative non-enhanced CT scans. The least absolute shrinkage and selection operator method was applied to select the most distinguishable features. Three logistic regression models were built to predict the EGFR mutation status by combining the CT semantic with clinicopathological characteristics, using the radiomic features alone, and by combining the radiomic and clinicopathological features. Receiver operating characteristic (ROC) curve analysis was performed using five-fold cross-validation and the mean area under the curve (AUC) values were calculated and compared between the models to obtain the optimal model for predicting EGFR mutation. Furthermore, radiomic nomograms were constructed to demonstrate the performance of the model.

In total, 1025 radiomic features were extracted and reduced to 13 features as the most important predictors to build the radiomic signature. The combined radiomic and clinicopathological features model was developed based on the radiomic signature, sex, smoking, vascular infiltration, and pathohistological type. The AUC was 0.90 $\pm$ 0.02 for the training, 0.88 $\pm$ 0.11 for the verification, and 0.894 for the test dataset. This model was superior to the other prediction models that used the combined CT semantic and clinicopathological features (AUC for the test dataset: 0.768) and radiomic features alone (AUC for the test dataset: 0.837).

The prediction model built by radiomic biomarkers and clinicopathological features, including the radiomic signature, sex, smoking, vascular infiltration, and pathological type, outperformed the other two models and could effectively predict the EGFR mutation status in patients with peripheral lung adenocarcinoma. The radiomic nomogram of this model is expected to become an effective biomarker for patients with lung adenocarcinoma requiring adjuvant targeted treatment.

Introduction

Lung cancer accounts for 13% of the global cancer incidence and is the leading cause of cancer-related death (Halpenny *et al* 2014, Siegel *et al* 2019). According to the histological type, it is divided into non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC accounts for 85 − 90% of lung cancer cases. Adenocarcinoma is the most common pathohistological subtype of NSCLC (Ganeshan *et al* 2012, Antonicelli *et al* 2013).

The introduction and progress of molecular targeted therapy have revolutionized NSCLC treatment, with significantly better selective tumor control than that in traditional chemotherapy and fewer toxic side effects (Nishino *et al* 2014, Ozkan *et al* 2015). Epidermal growth factor receptor (EGFR) is a transmembrane receptor tyrosine kinase involved in signaling pathways that regulate cell proliferation, apoptosis, angiogenesis, and invasion (Travis 2011, Nishino *et al* 2011). Small-molecule tyrosine kinase inhibitors (TKIs) against EGFRs were the first targeted drugs for NSCLC treatment. The EGFR mutation status is an important predictor of EGFR-TKI therapy efficacy in patients with NSCLC (Sugano *et al* 2011). Namely, the response rate to EGFR-TKIs in patients with EGFR mutations (60%–80%) is higher than that in patients with wild-type EGFRs or unknown mutations (10%–20%) (Riely *et al* 2006). Therefore, it is vital to identify the EGFR mutation status before treatment.

Detection of the EGFR mutation status is based on surgical specimen, biopsy sample, or hematological examination, all of which are invasive modalities. Additionally, due to the heterogeneity of tumors, the positive rate of EGFR mutation detection may vary across different tissue samples of the same patient. A further disadvantage is the high cost. Therefore, a noninvasive method with greater sensitivity and lower cost is necessary for detection of the EGFR mutation status in patients who would benefit from EGFR-TKI therapy.

Presently, multi-slice spiral computed tomography (MSCT) is a commonly used noninvasive examination method to analyze the tumor morphology and examine the correlation among the clinicopathological characteristics, CT imaging manifestations, and EGFR mutations in primary lung adenocarcinoma. Previous studies have shown that the EGFR mutation status is associated with many factors, such as the smoking status, pathohistological subtype, sex, and ethnicity (Russell *et al* 2013, Shi *et al* 2014). Recent studies have also shown that EGFR mutation is associated with the ground-glass opacity (GGO) (Lee *et al* 2013, Usuda *et al* 2014, Yang *et al* 2015). However, these studies have the limitation of predicting EGFR mutations through traditional, univariate analysis of CT or clinical features.

Radiomics is an emerging calculation method used for extracting all information contained in radiographic images for comprehensive systematic analysis. More precisely, radiomics is the use of automated algorithms to extract a large amount of features from the region of interest (ROI) of the image and further extract and strip the large-scale information through the statistical information and data mining methods to obtain key information that ultimately help the auxiliary diagnosis, classification, or grading of the disease (Kumar *et al* 2012, Lambin *et al* 2012, Parmar *et al* 2014, Gillies *et al* 2016). In recent years, some studies have found that radiomics is more meaningful in terms of tumor analysis and treatment than traditional clinical data (Leijenaar *et al* 2013, Kuo and Jamshidi 2014, Coroller *et al* 2015, Antunes *et al* 2016). Therefore, a logistic regression (LR) model with multivariate CT features and clinicopathological data was established in this study to predict the EGFR mutation status in peripheral lung adenocarcinoma.

Methods

Ethical considerations

This study was approved by our institutional review board.

Patient selection

We retrospectively reviewed the medical records of patients who underwent surgical resection for primary lung adenocarcinoma at our center from September 2016 to July 2018.

The inclusion criteria were as follows: (a) classification of the lung adenocarcinoma pathohistological subtype according to the 2015 World Health Organization (WHO) classification of lung cancer; (b) examination of the EGFR mutation status using amplification refractory mutation system-polymerase chain reaction (ARMS-PCR); (c) CT examination of the entire thorax using the same CT machine with the same slice thickness (1 mm) within 4 weeks before surgery; and (d) assay of carcinoembryonic antigen (CEA) before surgery. The exclusion criteria were as follows: (a) history of chemotherapy, radiotherapy, or extrathoracic metastases prior to undergoing CT examination; (b) insufficient or poor-quality tissue for molecular analyses; and (c) incomplete data.

The CT images and following clinicopathological data of all patients were collected for the analysis: age, sex, smoking history, pathohistological subtype, and EGFR mutation status.

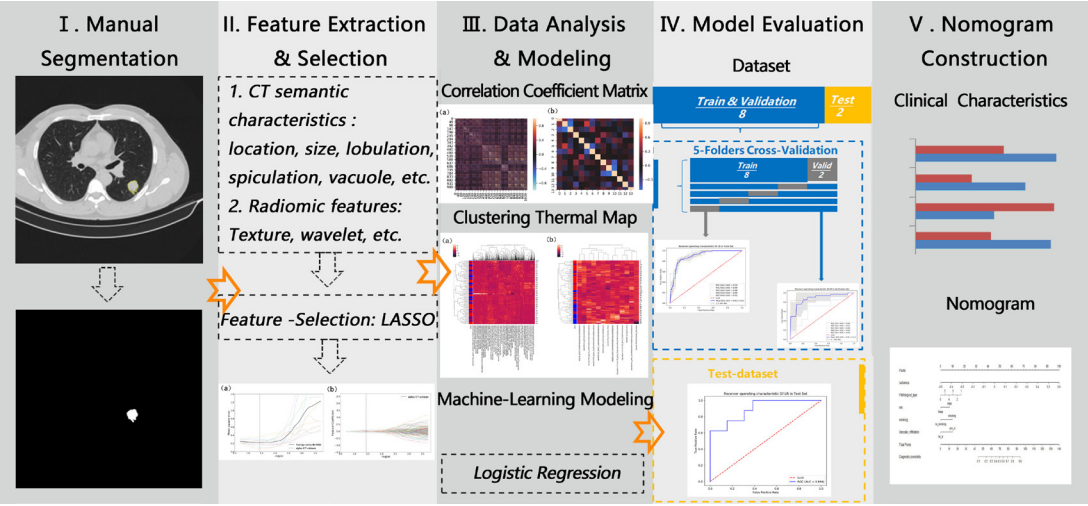

Figure 1. The workflow in this study.

Histopathological evaluation and molecular analysis

All resected specimens were formalin-fixed and stained with hematoxylin and eosin in accordance with the routine procedure in our hospital. The pathohistological subtype of the lung adenocarcinoma was classified according to the 2015 WHO classification of lung cancer. Paraffin specimens of tumor tissue were evaluated by two pathologists for the criterion of containing at least 50% tumor cells. The EGFR mutation status was examined by ARMS-PCR. All procedures were performed according to the manufacturer's protocol.

CT image acquisition and segmentation

All patients underwent non-enhanced CT examination of the entire thorax using a multidetector CT system (64-Slice; Siemens, Germany). In our study, the CT images were acquired using normalized protocols. The CT scan parameters were as follows: tube voltage, 120 kV; automatic tube current modulation, 35–90 mAs; pitch, 0.9; field of view, 180 mm × 180 mm; matrix, 512 × 512; reconstructed slice thickness and slice increment, both 1 mm. All CT images were exported in Digital Imaging and Communication in Medicine format for segmentation and image feature extraction.

Figure 1 shows the workflow of this study. Axial CT images were selected for analysis. The ROIs were delineated manually in a blind fashion by two highly-qualified radiologists in thoracic CT interpretation and were manually segmented using MicroDicom software.

Feature extraction and selection

In this study, semantic and radiomic features were extracted from the CT images. The two radiologists independently reviewed the preoperative thoracic CT images on the Picture Archiving and Communication System without knowing the patients' EGFR mutation status and evaluated the CT semantic features, including the type of lesion (solid type, pure or mixed GGO) and general tumor semantic features (such as location, size, lobulation, spiculation, and vacuolization).

The experiment was performed on a PC using a Windows 10 64-bit operating system with an Intel i7 CPU, 16GB RAM. The construction environment of the machine learning model was Python 3.6.1. The least absolute shrinkage and selection operator (LASSO) method was used for radiomic feature extraction and dimensionality reduction. The correlation coefficient of matrix thermal map and unsupervised clustering thermal map was calculated to judge the dimension reduction efficiency. Radiomic features were extracted from the ROIs using the pyRadiomics (2.0.1) package (van Griethuysen *et al* 2017) and included the tumor shape, intensity, texture, and wavelet features.

We also analyzed the interobserver reliability and intraobserver reproducibility of the two radiologists for the whole dataset in a blinded fashion.

Establishment of the prediction model and nomogram construction

Three LR models for prediction of EGFR mutation were established in the platform: (1) CT semantic features combined with clinicopathological features model, (2) radiomic biomarkers model, and (3) radiomic biomarkers combined with clinicopathological features model. A nomogram based on multivariate logistic analysis was implemented in R language to indicate the potential for EGFR mutation individually.

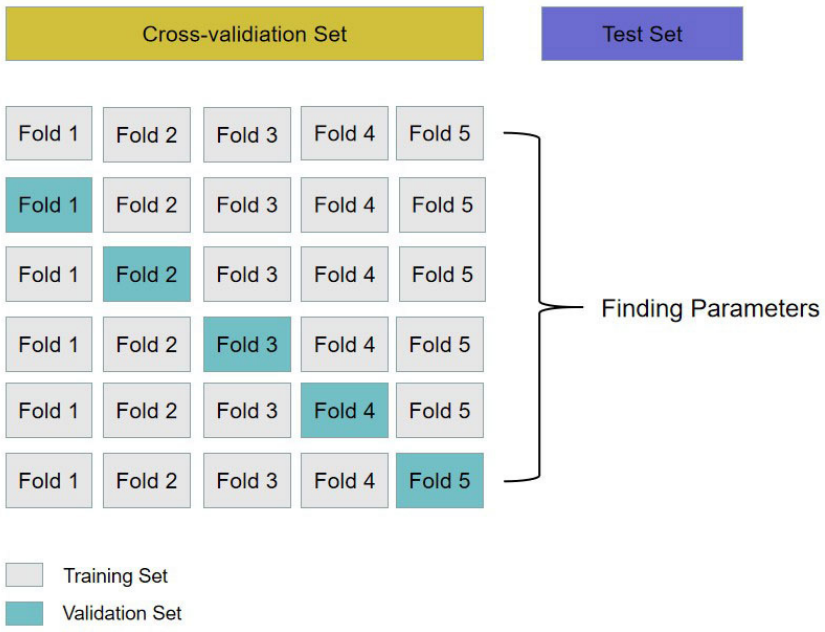

Figure 2. Datasets of the study.

The data were divided in two datasets: cross-validiation and test datasets, with a 8:2 proportion. Then the patients of cross-validiation dataset were evaluated by five-fold cross verification method. The first step is to randomly divide the original data into five copies without repeated samples. In the second step, one of them is selected as the verification dataset, and the remaining four are used as the training dataset. In the third step, the second step is repeated five times, so that each subset can be used as a verification and a training dataset. The fourth step is to calculate the average values of the test results of the five groups as the estimation of model accuracy and the performance index of the model under the current five-fold cross-validation. Finally, the test dataset is used to test the performance of the trained model. The training dataset was used to train the model, the validation dataset was used to adjust the complex parameters of the model, and the test dataset was used to test the performance of the model (figure 2).

Statistical analysis

Statistical analysis of the measurement and count data was performed using SPSS 24.0. Values of $P < 0.05$ were considered statistically significant. Receiver operating characteristic (ROC) curve analysis was performed and the mean area under the curve (AUC) value was calculated using five-fold cross-validation to determine the prediction ability of the LR models. Additionally, the prediction ability of the three models was compared to obtain the optimal model to predict EGFR mutation.

An independent samples t-test was used to evaluate the differences between the features generated by reader 1 and those by reader 2 (interobserver reliability), as well as the differences between the twice-generated features by reader 1 (intraobserver reproducibility). Inter- and intraclass correlation coefficients (ICCs) were used to evaluate the agreement of feature extraction. A good agreement was reached when the ICC was greater than 0.75.

Results

Clinicopathological data analysis

Of the total 824 patients whose records were reviewed, 104 patients were included in the analysis (figure 3). The statistical differences in EGFR mutation were detected between the cross-validation and test datasets in terms of age, sex, smoking status, CEA level, vascular infiltration, visceral pleural infiltration, lymph node metastasis, pathohistological subtype, and pathological stage. In this study, the rate of EGFR mutation was significantly higher in female than in male patients ($P = 0.015$) and in nonsmokers than in smokers ($P < 0.001$). The most common pathohistological subtypes of lung adenocarcinoma were the acinar and lepidic subtypes. The pathohistological subtypes of lung adenocarcinoma and the vascular infiltration status were correlated with EGFR mutation ($P = 0.024$ and $P = 0.030$, respectively). Therefore, the sex, smoking status, pathohistological subtype, and vascular infiltration status were used for the model and nomogram establishment. The results of the correlation analysis between table 1.

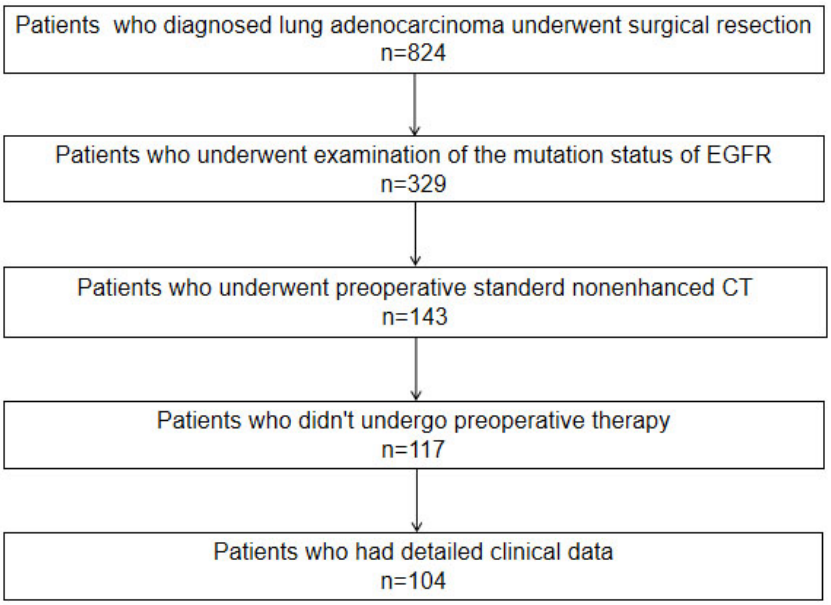

Figure 3. Recruitment criteria for patients.

CT semantic and radiomic features extraction and selection

CT semantic features extraction and selection

In this study, the CT semantic and clinicopathological features of the patients were represented as 45 categorical variables. Redundant features were ruled out by the LASSO dimensional reduction model, retaining six features, including sex, age, visceral pleural infiltration, papillary pathohistological subtype, spiculation, and tumor necrosis, as shown in figure 4.

Radiomic features extraction and selection

Using the LASSO regression model (figure 5), the original 1025-dimensional radiomic features were shrunk to 13 nonzero radiomic features (table 2).

Radiomic feature verification

The unsupervised clustering thermal map and correlation coefficient matrix thermal map were used to preliminarily verify the dimensionality reduction of the radiomic features and evaluate the dimensionality reduction efficiency.

The unsupervised clustering thermal map was derived through clustering, which was an unsupervised learning method, of the 1025 pre-reduction radiomic features and the 13 phylogenetic post-reduction features. The clustering performance of the pre-reduction radiomic features was lower than that of the post-reduction features, indicating that the 13 radiomic features obtained with the LASSO dimensional reduction model are superior (figure 6).

The correlation coefficient matrix thermal map was drawn for the 1025 classic features and the 13 classic features of the experiment. Values close to '0' indicated a lack of correlation, those closer to '1' indicated a positive correlation, and those closer to ' $-$ 1' indicated a negative correlation. The results showed that the correlation coefficient matrix after dimensionality reduction retained the distribution characteristics of that before dimensionality reduction and filtered out the redundant features (figure 7).

Development of the prediction models and ROC curve analysis

Development of the CT semantic features combined with the clinicopathological features model

This LR model was constructed according to the six features retained in the training dataset. Five-fold cross-validation was used in this experiment. The ROC curves of the training, validation, and test datasets for the semantic modeling are shown in figure 8. The important parameters were determined by the five-fold cross-validation, and the parameters of the LR classifier were as follow: solver $=$ 'liblinear', tol $= 0.0001$, $C = 1.0$. The AUC of the prediction model was 0.78 ± 0.02 in the training, 0.69 ± 0.10 in the verification, 0.769 in the cross-validation and 0.768 in the test dataset.

Table 1. Correlation analysis between clinicopathological characteristics and EGFR mutations in patients with peripheral lung adenocarcinoma.

	Cross-validiation set			Test set		
Clinicopathological Characteristics	EGFR +	EGFR −	*P*	EGFR +	EGFR −	*P*
No. of patients	51	32	—	13	8	—
Age, mean ± STD	58.00 ± 8.10	58.53 ± 10.59	0.809	56.77 ± 11.13	55.88 ± 14.61	0.875
Gender						
Male	15 (29.4%)	18 (56.3%)	0.015[a]	2 (15.4%)	5 (62.5%)	0.030[a]
Female	36 (70.6%)	14 (43.7%)		11 (84.6%)	3 (37.5%)	
Smoking status						
No	44 (86.3%)	16 (50.0%)	0.000[a]	11 (84.6%)	3 (37.5%)	0.030[a]
Yes	7 (13.7%)	16 (50.0%)		2 (15.4%)	5 (62.5%)	
CEA						
Normal	38 (74.5%)	19 (59.4%)	0.148	9 (69.2%)	4 (50.0%)	0.646
Abnormal	13 (25.5%)	13 (40.6%)		4 (30.8%)	4 (50.0%)	
Vascular infiltration						
No	46 (90.2%)	23 (71.9%)	0.030[a]	10 (76.9%)	2 (25.0%)	0.032[a]
Yes	5 (9.8%)	9 (28.1%)		3 (23.1%)	6 (75.0%)	
Visceral pleural infiltration						
No	18 (35.3%)	6 (18.8%)	0.106	3 (23.1%)	3 (37.5%)	0.631
Yes	33 (64.7%)	26 (81.2%)		10 (76.9%)	5 (62.5%)	
Lymph node metastasis						
No	42 (82.4%)	21 (65.6%)	0.083	9 (69.2%)	5 (62.5%)	1.000
Yes	9 (17.6%)	11 (34.4%)		4 (30.8%)	3 (37.5%)	
Histological subtype						
Solid	4 (7.8%)	6 (18.8%)	0.024[a]	0 (0.0%)	1 (12.5%)	0.025[a]
Papillary	4 (7.8%)	6 (18.8%)		2 (15.4%)	0 (0.0%)	
Micro-papillary	2 (3.9%)	2 (6.3%)		1 (7.7%)	3 (37.5%)	
Acinar	31 (60.8%)	14 (43.8%)		5 (38.5%)	3 (37.5%)	
Lepidic	10 (19.6%)	4 (12.5%)		5 (38.5%)	0 (0.0%)	
Infiltrating mucinous	0 (0.0%)	0 (0.0%)		0 (0.0%)	1 (12.5%)	
Stage						
I	37 (72.5%)	17 (53.1%)	0.060	9 (69.2%)	4 (50.0%)	0.316
II	8 (15.7%)	7 (21.9%)		0 (0.0%)	1 (12.5%)	
III	4 (7.8%)	8 (25.0%)		4 (30.8%)	3 (37.5%)	
IV	2 (3.9%)	0 (0.0%)		0 (0.0%)	0 (0.0%)	

[a] $P < 0.05$, the difference was statistically significant. There were statistically significant differences between the two groups in sex, smoking status, vascular infiltration, and pathohistological subtype.

Development of the radiomic features model

This LR model was established according to the 13-dimensional radiomic features of the training dataset through five-fold cross-validation. The established and determined important parameters of the LR classifier were as follows: solver = 'liblinear', tol = 0.0001, $C = 1.0$. As shown in the results (figure 9), the AUC of this prediction model was 0.92 ± 0.01 in the training, 0.84 ± 0.04 in the verification, 0.907 in the cross-validation, and 0.837 in the test dataset.

The prediction ability of the radiomic features LR model for EGFR mutation in the training and test datasets was further verified by constructing a box plot (figure 10). Because the average value is greatly affected by the extreme value, it is sometimes unreasonable to use it to measure the overall situation, whereas the median value is not affected by the extreme value. Thus, the median value was considered more suitable to represent the 1D radiomic features. In both datasets, the median difference between the two groups of data was large, indicating that the radiomic features LR model is better for the classification of negative-positive patients.

Development of the radiomic features combined with the clinicopathological features model

This LR model was established in the same manner by combining the 13 radiomic features with the clinicopathological features that were found to be associated with EGFR mutation: sex, smoking status, vascular infiltration status, and pathohistological subtype (figure 11). We mapped the 13 radiomics features to a single radiomics feature by equation

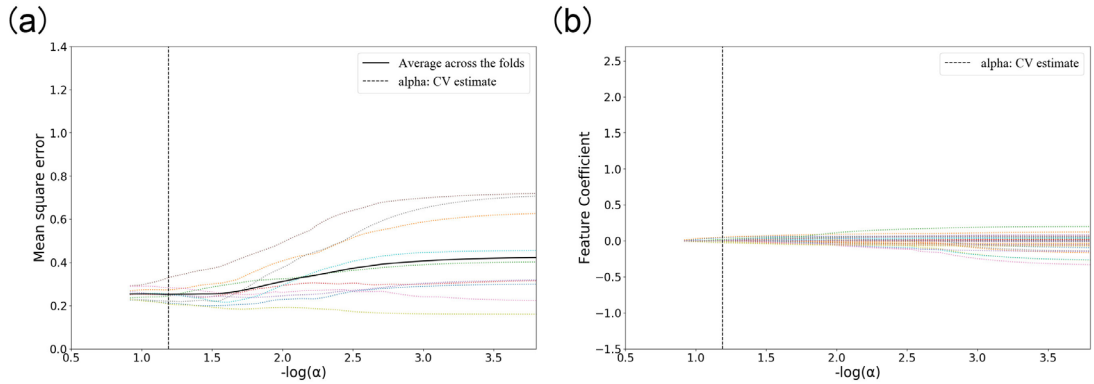

Figure 4. (a) α is the penalty term coefficient, and the ordinate is the mean square error. The ten dashed curves represent the relationship between the value of the negative logarithm of each alpha and the corresponding mean square error in the ten-fold cross-validation method. The solid line in the figure represents the mean value of the ten-fold cross-validation corresponding to the dashed curve, and the vertical dashed line represents the $(-\log(\alpha))$ value corresponding to the optimal α value when the mean square error is the smallest. The final value of α was 0.064. (b) The convergence graph of the characteristic coefficients of the screened features by the LASSO model ten-fold cross-validation method. Each curve in the figure represents the change of the characteristic coefficient of one of the 45 categorical variables with the α value, and the vertical dashed line indicates the $(-\log(\alpha))$ value corresponding to the optimal α value to obtain six features with nonzero characteristic coefficients.

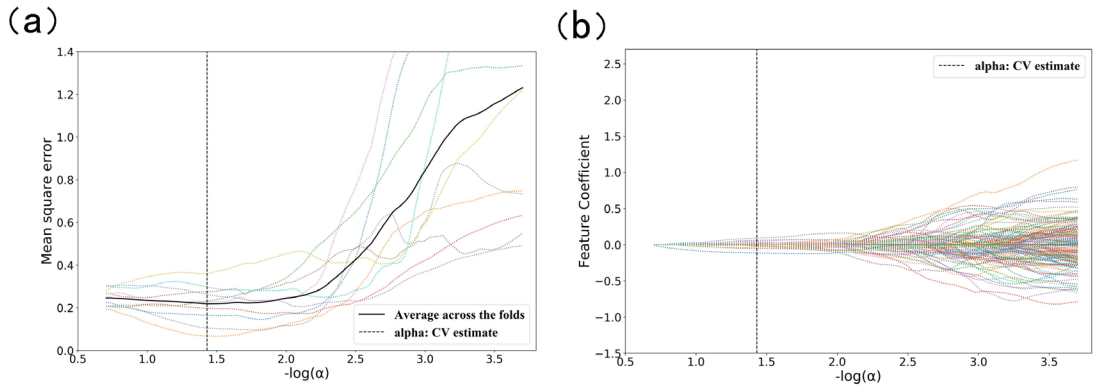

Figure 5. LASSO dimensionality reduction curve of radiomic features.

Table 2. Names and coefficients of non-zero radiomics features.

Name	Coefficient
Original_glszm_SmallAreaHighGrayLevelEmphasis	0.027 3974
Original_glszm_SmallAreaLowGrayLevelEmphasis	−0.117 795
Original_gldm_LowGrayLevelEmphasis	0.030 3573
Log-sigma-1-0 mm-3D_glcm_Cluster Prominence	0.028 5841
Log-sigma-3-0 mm-3D_gldm_DependenceNonUniformityNormalized	0.026 4742
Wavelet-LLL_glcm_InverseVariance	−0.083 8051
Wavelet-LLH_glcm_Imc2	0.063 1772
Wavelet-HLL_firstorder_Mean	0.049 0448
Wavelet-HLL_glszm_LowGrayLevelZoneEmphasis	0.032 4594
Wavelet-HLL_gldm_SmallDependenceHighGrayLevelEmphasis	−0.043 8279
Wavelet-HLH_glszm_SizeZoneNonUniformityNormalized	0.033 0170
Wavelet-HHH_firstorder_Skewness	−0.013 6832
Wavelet-HHH_glszm_Size Zone Non Uniformity Normalized	0.016 3165

$$R^* = \sum_i^n v_i * c_i$$

where n was the number of radiomics features, v_i was each radiomics feature value and c_i represent the corresponding coefficient of each image feature value. The important parameters of the LR classifier were

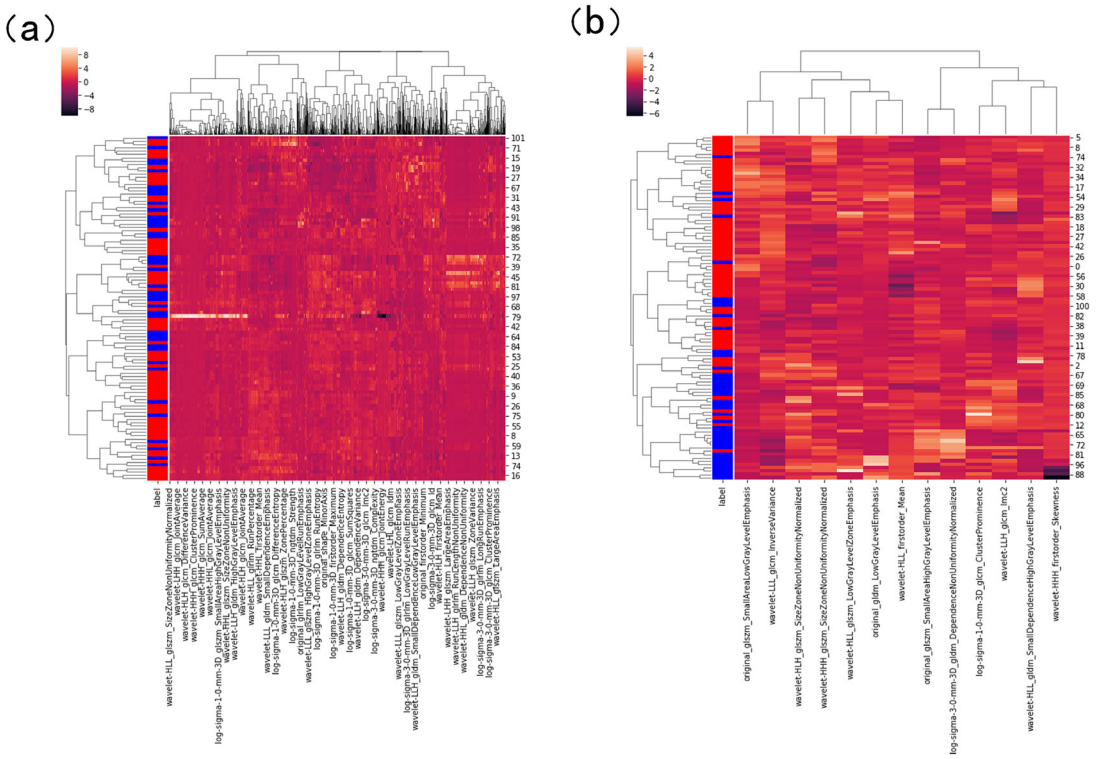

Figure 6. Unsupervised clustering thermal map before (a) and after (b) dimensionality reduction. The leftmost red color in the graph represents the category of EGFR mutations and the blue color represents the category of wild-type EGFR. The cluster distribution of the right graph is more concentrated, indicating that the 13 radiomic features obtained with the LASSO dimensional reduction model are superior.

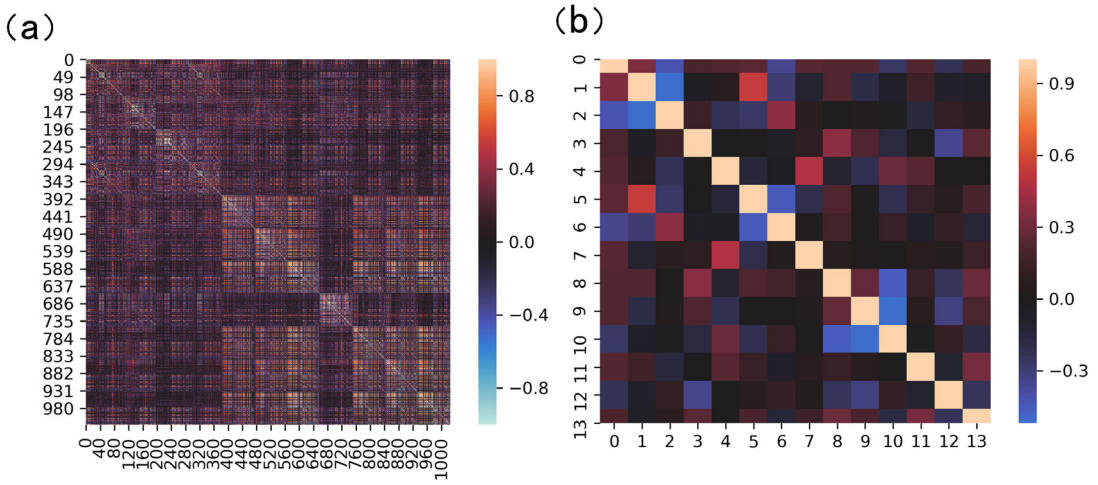

Figure 7. Correlation coefficient matrix thermal map before (a) and after (b) dimensionality reduction.

solver = 'liblinear', tol = 0.0001, $C = 1.0$. The results showed that the AUC of the prediction model was 0.90 ± 0.02 in the training, 0.88 ± 0.11 in the verification, 0.898 in the cross-validation and 0.894 in the test dataset.

Additionally, the sensitivity, specificity, positive predictive value, and negative predictive value were calculated for each model to demonstrate the predictive power (table 3).

Validation of the radiomic nomogram

We trained the above model and obtained the Rad-score formula of the radiomics as follows:

$$\text{Rad-score} = -0.360\,0158 + 1.511\,727\,63 * \text{radiomics} - 0.210\,275\,81 * \text{sex} + 0.354\,184\,95 * \text{smoking} + 0.485\,570\,62 * \text{Vascular\_infiltration} - 0.384\,636\,02 * \text{Pathological\_type}.$$

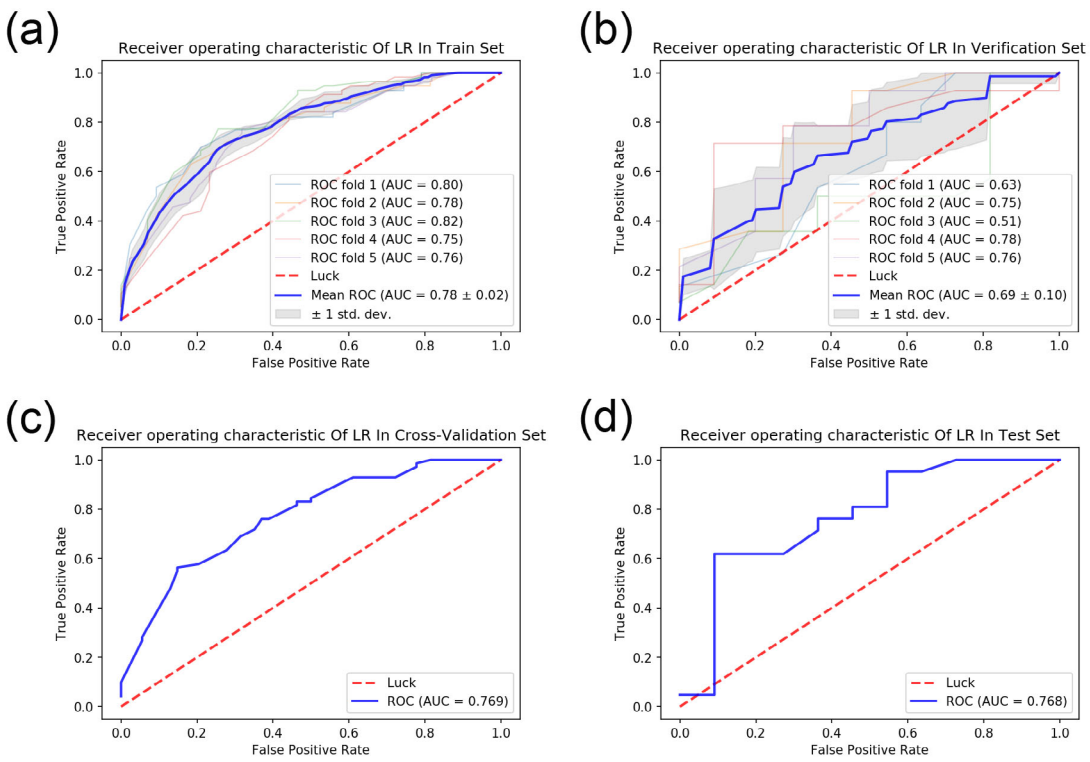

Figure 8. ROC curves of the training (a), verification (b), cross-validation (c), and test (d) datasets for the CT semantic model.

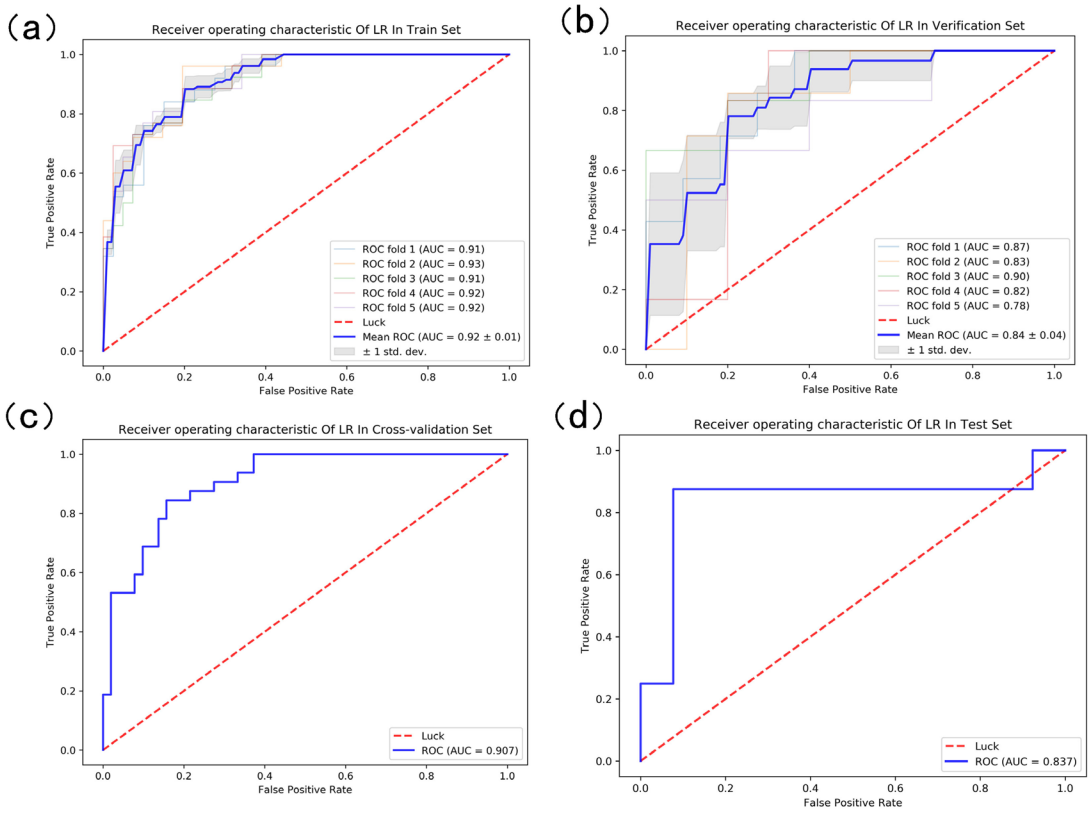

Figure 9. ROC curves of the training (a), verification (b), cross-validation (c), and test (d) datasets for the radiomic features model.

According to the Rad-score formula, the individualized EGFR mutation prediction model comprised the above predictors by the nomogram. The nomogram, built in the training dataset, represented the relationship between the radiomic and clinicopathological features (sex, smoking, vascular infiltration status, and pathohistological type) and visually showed the potential ability (figure 12). The corresponding Rad-scores of the different

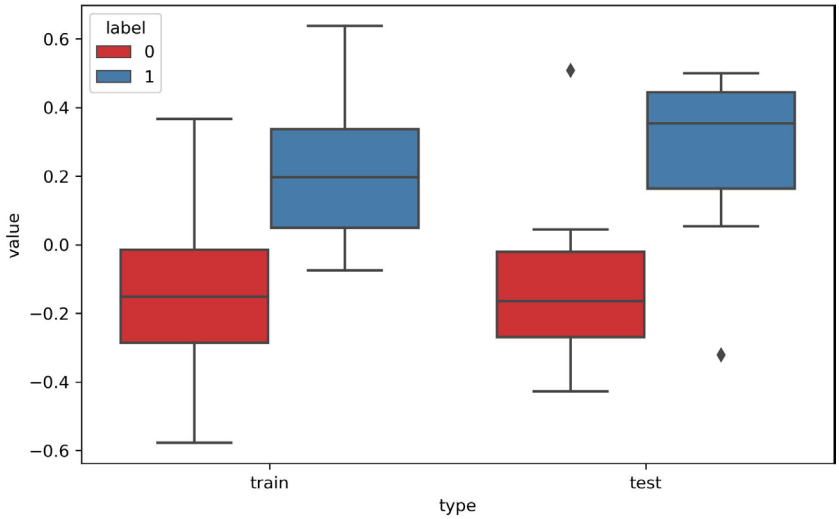

Figure 10. Box plot of EGFR mutation predicted by the radiomic features LR model. Blue color represents EGFR wild type and red color represents EGFR mutation type. The left side of the abscissa is the training set, and the right side is the test set. In the training set, the negative overall distribution is around −0.17 and the positive overall distribution is approximately 0.2, which is more distinguished. Similarly, the median difference between the two groups of data was large in the test set. Thus, the radiomic LR model is better for classification of negative-positive patients.

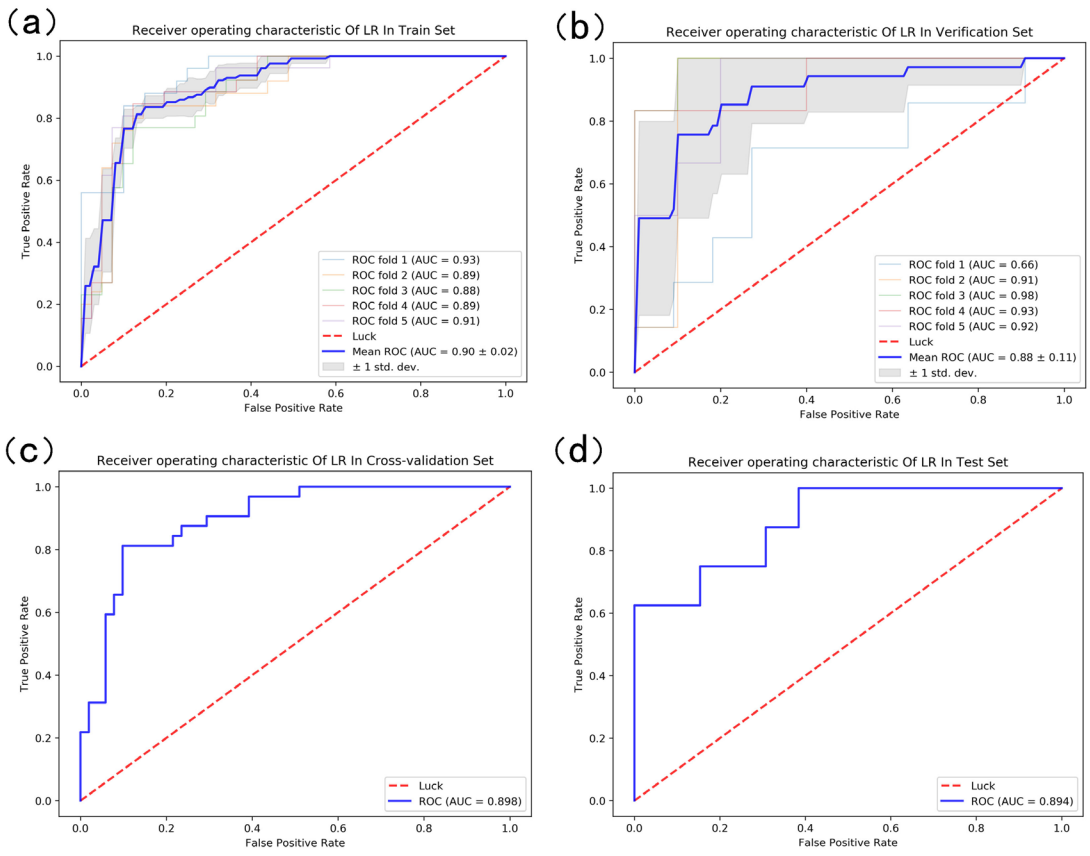

Figure 11. ROC curves of the training, verification, cross-validation, and test datasets for the radiomic features combined with clinicopathological features model.

features were defined, and the total Rad-score was obtained by accumulating all points. Finally, the rate of EGFR mutation, namely, the diagnostic possibility, was obtained by the total Rad-score. An EGFR mutation rate of 50% corresponded to a Rad-score of 69.

Table 3. Accuracy of EGFR mutation predicted by the three models in the training, verification, cross-validation, and test datasets.

Model	Cohort	Sensitivity	Specificity	Positive predictive values	Negative predictive values
CT Semantic And Clinicopathological Characteristics model	Training cohort of cross-validation	0.95 ± 0.01	0.22 ± 0.03	0.65 ± 0.01	0.74 ± 0.05
	Validation cohort of cross-validation	0.95 ± 0.05	0.23 ± 0.01	0.66 ± 0.03	0.77 ± 0.20
	Cross-validation cohort	0.95	0.22	0.65	0.73
	Test cohort	1.00	0.19	0.55	1.00
Radiomics model	Training cohort of cross-validation	0.73 ± 0.01	0.92 ± 0.01	0.85 ± 0.02	0.84 ± 0.00
	Validation cohort of cross-validation	0.59 ± 0.21	0.82 ± 0.04	0.65 ± 0.16	0.77 ± 0.08
	Cross-validation cohort	0.69	0.88	0.79	0.82
	Test cohort	0.75	0.92	0.86	0.86
Radiomics and clinicopathological characteristics model	Training cohort of cross-validation	0.81 ± 0.04	0.87 ± 0.01	0.79 ± 0.01	0.88 ± 0.02
	Validation cohort of cross-validation	0.80 ± 0.26	0.80 ± 0.16	0.72 ± 0.16	0.89 ± 0.13
	Cross-validation cohort	0.81	0.86	0.79	0.88
	Test cohort	0.75	0.85	0.75	0.85

Discussion

EGFR-TKIs are the most effective first-line therapeutic modality for lung adenocarcinoma with EGFR mutation. Compared with traditional chemotherapy, it results in a longer progression-free survival and a higher quality of life. Therefore, it is extremely important to identify the EGFR mutation status. Efficient noninvasive detection of the EGFR mutation status is necessary for patients preparing to accept EGFR-TKIs. In this study, we extracted and combined the clinicopathological, CT semantic, and CT radiomic features of patients with lung adenocarcinoma using a series of rigorous statistical verification analyses to assess the EGFR mutation status.

As the most frequently used method for lung cancer screening, chest CT could provide various imaging features, including the tumor location, quantity, size, density, calcifications, focal necrosis, cavitation, vacuolization, spiculation, lobulation, pleural indentation, and pleural effusion. Radiomics enables extraction and analysis of these features to obtain information that can aid in the diagnosis, classification, or grading of a disease, as well as in treatment response prediction. In our study, three LR prediction models were established: CT semantic features combined with clinicopathological features model, radiomic features model, and radiomic features combined with clinicopathological features model.

The CT semantic features combined with clinicopathological features model was developed, which comprised six features, including sex, age, visceral pleural infiltration, papillary pathohistological subtype, spiculation, and tumor necrosis. Because the CT semantic features were low-dimensional, EGFR mutation could not be well predicted in this feature space. Therefore, high-dimensional radiomic features were used for modeling to obtain better prediction results. Thirteen potential predictors were implemented to develop the radiomic features model. Subsequently, the radiomic features combined with clinicopathological features model was established by combining the radiomic features with the clinicopathological features that were found to be associated with EGFR mutation, including the sex, smoking status, vascular infiltration, and pathohistological subtype. In the test dataset, the AUC values of the combined radiomic and clinicopathological features model were higher than those of the LR model with radiomic features alone, indicating that the model with combined radiomic and clinicopathological features could better predict EGFR mutation.

Previous studies have used radiomics to predict lung cancer gene mutations. Sacconi *et al* (2017) analyzed the correlation of quantitative texture radiomic features with EGFR mutation and survival rates in 68 patients with lung cancer. They found that the mean value ($P = 0.0001$), standard deviation ($P = 0.0001$), and skewness ($P = 0.0459$) were significantly correlated with EGFR mutation, while entropy was the only variable correlated with mortality ($r = 0.2708$, $P = 0.0329$). Rios Velazquez *et al* (2017) studied the relationship between the radiomic features of lung adenocarcinoma and EGFR mutation. Their results showed that the AUC value of EGFR mutation predicted by the radiomic features alone was 0.69; thus, they combined this model with the clinical features model (AUC = 0.70) to improve the prediction accuracy (AUC = 0.75). Liu *et al* (2016) studied EGFR

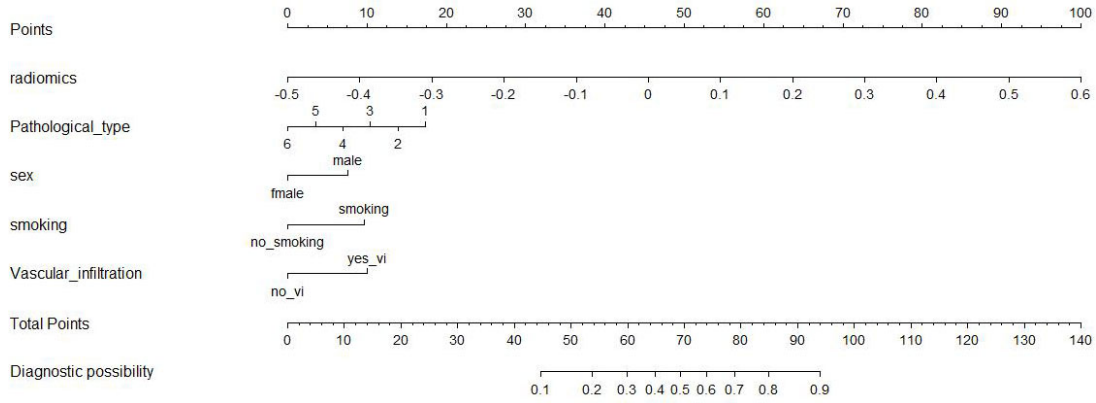

Figure 12. Radiomic nomogram validation. The nomogram was built in the training dataset with the radiomic features, sex, smoking, vascular infiltration status, and pathohistological type (No. 1-6 represent the solid, papillary, micropapillary, acinar, lepidic, and infiltrating mucinous subtype, respectively). An EGFR mutation rate of 50% corresponded to a Rad-score of 69.

mutation in patients with peripheral lung adenocarcinoma using the radiomics method and obtained 219 radiomic features, among which, five features combined with clinical features could successfully predict EGFR mutation with an AUC value of 0.709. Our findings are consistent with the results of these studies: combined radiomic and clinical features could successfully predict EGFR mutations. However, our findings are significant in that a much greater number of radiomic features were extracted (1025), the AUC value was significantly higher than those in the previous studies, and a nomogram was created for individualized prediction.

In this study, five-fold cross verification was used by which the sample data were divided into training, validation, and test datasets. Through five-fold cross-validation, the results of five different training groups are averaged to reduce variance, thus the performance of the model is not so sensitive to data partition. Then the test dataset is used to test the performance of the trained model. In previous studies, the research model exclusively had training and verification datasets, with a ROC curve and an AUC value. When the training and verification datasets are randomly grouped again, the results will alter and be chosen as one of the highest AUC values in the final model, leading to certain bias and questioning the authenticity of the data.

There are some limitations to this study. First, the sample size was not substantial, and only 104 patients met the inclusion criteria. Second, the study was retrospective and included only Eastern Asian population, which limits the generalizability of the results. Therefore, future studies with a larger sample size should be performed and more Western population patients should participate to verify our findings and extend their generalizability. Moreover, the radiomic method can be combined with deep learning methods to improve the performance of the model.

Conclusion

This study revealed the correlation of clinicopathological and CT imaging features with the EGFR mutation status in patients with peripheral lung adenocarcinoma. The combined radiomic and clinicopathological features model, comprising the radiomic signature, sex, smoking status, vascular infiltration status, and pathohistological type, could effectively predict the EGFR mutation status. It was presented as a radiomic nomogram that is expected to become an effective biomarker for populations requiring adjuvant EGFR-TKI targeted treatment.

Acknowledgments

This work was supported by the National Health Commission of the People's Republic of China (Grant Nos. 131025000000170001); the Natural Science Foundation of Jilin Province (Grant Nos. 20180101038JC); the Science and Technology Development Plan of Jilin Province (Grant Nos. 20170622009JC); and the Provincial and School Joint Construction Project of Jilin University (Grant Nos. SXGJXX2017-8).

ORCID iDs

Xueyan Li https://orcid.org/0000-0002-9206-9480

References

Antonicelli A *et al* 2013 EGFR-targeted therapy for non-small cell lung cancer: focus on EGFR oncogenic mutation *Int. J. Med. Sci.* **10** 320–30

Antunes J, Viswanath S, Rusu M, Valls L, Hoimes C, Avril N and Madabhushi A 2016 Radiomics analysison FLT-PET/MRI for characterization of early treatment response in renal cell carcinoma: a proof-of-concept study *Transl. Oncol.* **9** 155–62

Coroller T P *et al* 2015 CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma *Radiother. Oncol.* **114** 345–50

Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S and Miles K 2012 Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival *Eur. Radiol.* **22** 796–802

Gillies R J, Kinahan P E and Hricak H 2016 Radiomics: Images are more than pictures, they are data *Radiology* **278** 563–77

Halpenny D F, Riely G J, Hayes S, Yu H, Zheng J, Moskowitz C S and Ginsberg M S 2014 Are there imaging characteristics associated with lung adenocarcinomas harboring ALK rearrangements? *Lung Cancer* **86** 190–4

Kumar V *et al* 2012 Radiomics: the process and the challenges *Magn. Reson. Imaging* **30** 1234–48

Kuo M D and Jamshidi N 2014 Behind the numbers: decoding molecular phenotypes with radiogenomics–guiding principles and technical considerations *Radiology* **270** 320–5

Lambin P *et al* 2012 Radiomics: extracting more information from medical images using advanced feature analysis *Eur. J. Cancer* **48** 441–6

Lee H J, Kim Y T, Kang C H, Zhao B, Tan Y, Schwartz L H, Persigehl T, Jeon Y K and Chung D H 2013 Epidermal growth factor receptor mutation in lung adenocarcinomas: relationship with CT characteristics and histologic subtypes *Radiology* **268** 254–64

Leijenaar R T H *et al* 2013 Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability *Acta. Oncol.* **52** 1391–7

Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia A L, Stringfield O, Ye Z and Gillies R J 2016 Radiomic features are associated with EGFR mutation status in lung adenocarcinomas *Clin. Lung Cancer* **17** 441–8

Nishino M, Hatabu H, Johnson B E and McLoud T C 2014 State of the art:response assessment in lung cancer in the era of genomic medicine *Radiology* **271** 6–27

Nishino M, Jackman D M, Hatabu H, Jänne P A, Johnson B E and Van den Abbeele A D 2011 Imaging of lung cancer in the era of molecular medicine *Acad. Radiol.* **18** 424–36

Ozkan E *et al* 2015 CT Gray-level texture analysis as a quantitative imaging biomarker of epidermal growth factor receptor mutation status in adenocarcinoma of the lung *Am. J. Roentgenol.* **205** 1016–25

Parmar C *et al* 2014 Robust radiomics feature quantification using semiautomatic volumetric segmentation *PLoS One* **9** e102107

Riely G J, Pao W, Pham D, Li A R, Rizvi N, Venkatraman E S, Zakowski M F, Kris M G, Ladanyi M and Miller V A 2006 Clinical course of patients with non–small cell lung cancer and epidermal growth factor receptor exon 19 and exon 21 mutations treated with gefitinib or erlotinib *Clin. Cancer Res.* **12** 839–44

Rios Velazquez E *et al* 2017 Somatic mutations drive distinct imaging phenotypes in lung cancer *Cancer Res.* **77** 3922–30

Russell P A *et al* 2013 Correlation of mutation status and survival with predominant histologic subtype according to the new IASLC/ATS/ERS lung adenocarcinoma classification in stage III (N2) patients *J. Thorac. Oncol.* **8** 461–8

Sacconi B *et al* 2017 Analysis of CT features and quantitative texture analysis in patients with lung adenocarcinoma: a correlation with EGFR mutations and survival rates *Clin. Radiol.* **72** 443–50

Shi Y, Au J S, Thongprasert S, Srinivasan S, Tsai C M, Khoa M T, Heeroma K, Itoh Y, Cornelio G and Yang P C 2014 A prospective molecular epidemiology study of EGFR mutations in Asian patients with advanced non–small-cell lung cancer of adenocarcinoma histology (PIONEER) *J. Thorac. Oncol.* **9** 154–62

Siegel R L, Miller K D and Jemal A 2019 Cancer statistics *CA Cancer J. Clin.* **69** 7–34

Sugano M, Shimizu K, Nakano T, Kakegawa S, Miyamae Y, Kaira K, Araki T, Kamiyoshihara M, Kawashima O and Takeyoshi I 2011 Correlation between computed tomography findings and epidermal growth factor receptor and Kras gene mutations in patients with pulmonary adenocarcinoma *Oncol. Rep.* **26** 1205–11

Travis W D 2011 Pathology of lung cancer *Clin. Chest Med.* **32** 669–92

Usuda K *et al* 2014 Relationships between EGFR mutation status of lung cancer and preoperative factors - are they predictive? *Asian Pac. J. Cancer Prev.* **15** 657–62

van Griethuysen J J M, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G H, Fillion-Robin J C, Pieper S and Aerts H J W L 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7

Yang Y, Yang Y, Zhou X, Song X, Liu M, He W, Wang H, Wu C, Fei K and Jiang G 2015 EGFR L858R mutation is associated with lung adenocarcinoma patients with dominant ground-glass opacity *Lung Cancer* **87** 272–7