Development and Validation of a Concise Prediction Scoring System for Asian Lung Cancer Patients with EGFR Mutation Before Treatment

Wenting An, MM¹, Wei Fan, MM², Feiyang Zhong, MM¹, Binchen Wang, MM¹, Shan Wang, MM¹, Tian Gan, MM¹, Sufang Tian, MD², and Meiyan Liao, MD¹

Abstract
Purpose We aimed to determine the epidermal growth factor receptor (EGFR) genetic profile of lung cancer in Asians, and develop and validate a non-invasive prediction scoring system for EGFR mutation before treatment. Methods This was a single-center retrospective cohort study using data of patients with lung cancer who underwent EGFR detection (n = 1450) from December 2014 to October 2020. Independent predictors were filtered using univariate and multivariate logistic regression analyses. According to the weight of each factor, a prediction scoring system for EGFR mutation was constructed. The model was internally validated using bootstrapping techniques and temporally validated using prospectively collected data (n = 210) between November 2020 and June 2021. Results In 1450 patients with lung cancer, 723 single mutations and 51 compound mutations were observed in EGFR. Thirty-nine cases had two or more synchronous gene mutations. We developed a scoring system according to the independent clinical predictors and stratified patients into risk groups according to their scores: low-risk (score <4), moderate-risk (score 4-8), and high-risk (score >8) groups. The C-statistics of the scoring system model was 0.754 (95% CI 0.729-0.778). The factors in the validation group were introduced into the prediction model to test the predictive power of the model. The results showed that the C-statistics was 0.710 (95% CI 0.638-0.782). The Hosmer–Lemeshow goodness-of-fit showed that $\chi^2 = 6.733$, P = 0.566. Conclusions The scoring system constructed in our study may be a non-invasive tool to initially predict the EGFR mutation status for those who are not available for gene detection in clinical practice.

Keywords
epidermal growth factor receptor, lung cancer, predictive model, scoring system

Abbreviations:
AFP, alpha fetoprotein; AIS, adenocarcinoma in situ; ARMS-PCR, amplification refractory mutation system-polymerase chain reaction; Braf, B-Raf proto-oncogene, serine/threonine kinase; CA, carbohydrate antigen; CEA, carcinoembryonic antigen; CI, confidence interval; CK, cytokeratin; CT, computerized tomography; CTNNB1, catenin beta 1; EGFR, epidermal growth factor receptor; EML4, echinoderm microtubule-associated protein-like 4; ERBB2, erb-b2 receptor tyrosine kinase 2; ERBB4, erb-b2 receptor tyrosine kinase 4; FBXW7, F-box and WD repeat domain containing 7; FERR, ferritin; FGFR3, fibroblast growth factor receptor 3; GGO, ground-glass opacity; HRAS, HRas proto-oncogene; HRCT, high resolution computed tomography;

¹ Zhongnan Hospital of Wuhan University, Wuhan, China
² Zhongnan Hospital of Wuhan University, Wuhan, China

Corresponding Authors:
Meiyan Liao, Department of Radiology, Zhongnan Hospital of Wuhan University, No.169 Donghu Road, Wuchang District, Wuhan City 430071, Hubei Province, China.
Email: liaomy@whu.edu.cn

Sufang Tian, Department of Pathology, Zhongnan Hospital of Wuhan University, No.169 Donghu Road, Wuchang District, Wuhan City 430071, Hubei Province, China.
Email: sftian@whu.edu.cn

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).
KRAS, KRAS proto-oncogene; MET, mesenchymal to epithelial transition factor; NGS, next-generation sequencing; NSCLC, non-small cell lung cancer; N Ras, NRAS proto-oncogene; NSE, neuron-specific enolase; OR, odds ratio; PIK3CA, phosphatidylinositol-4, 5-bisphosphate 3-kinase catalytic subunit alpha; PTEN, phosphatase and tensin homolog; ROC, receiver operating characteristic curve; ROS1, c-ros oncogene 1; SCCA, squamous cell carcinoma antigen; CYFRA21-1, soluble fragment of cytokeratin 19; STK11, serine/threonine kinase 11; TKI, tyrosine kinase inhibitor; TP53, tumor protein p53; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; TSC1, TSC complex subunit 1; TTF-1, thyroid transcription factor-1; UICC/AJCC, International Union Against Cancer and the American Joint Committee on Cancer.

**Introduction**

Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for 22%–23% of all cancer-related deaths.\(^1\) During recent years, with the rapid development of precision medicine and tumor molecular biology, targeted therapy has become an important treatment for lung cancer following the traditional approaches of surgery, radiotherapy, and chemotherapy. Considerably high positive response rate and safety have made individualized and refined treatment possible. Epidermal growth factor receptor (EGFR)-tyrosine kinase inhibitors (TKIs) such as gefitinib, osimertinib, and erlotinib have increased the overall survival of patients expressing EGFR.\(^2\) EGFR mutation is the most common genetic mutation related to lung cancer, with a mutational frequency of approximately 46%–58% in China.\(^3\)\(^–\)\(^5\)

Currently, the sources of tumor material for gene detection include tumor tissue specimens, cytology specimens, and serum specimens. Paraffin-embedded tumor tissue specimens have conventionally been the main source and still account for most diagnostic samples in clinical practice. Cytology specimens have been shown to be an adequate alternative source when tissue samples are not available or contain insufficient amounts of tumor DNA, and their use has increased over recent years.\(^6\) Serum specimens, known as liquid biopsies, are not popular in clinical practice owing to the lack of standardized techniques, limited coverage of hotspot mutations, lack of sensitivity, and insufficient clinical validation.\(^7\) For patients with advanced lung cancer, biopsy specimens are the only available histological evidence. However, tumor heterogeneity is a major challenge, as gene information from local tumor tissue might not accurately reflect the whole genetic profile. In addition, these specimens may be difficult to obtain from some patients with advanced age, poor pulmonary function, or poor coagulation. Furthermore, in some remote areas in Asia, some patients cannot afford the cost of hospitalization, and primary hospitals lack gene detection technology. Thus, it is necessary to establish a non-invasive and convenient method to initially predict the EGFR mutation status and evaluate the clinical treatment efficacy.

To date, there have been various small-scale studies revealing the relationship between some clinicopathological characteristics and EGFR mutation.\(^8\)\(^–\)\(^9\) A few studies have also established the prediction scoring system for EGFR mutation but have low credibility because of the use of too few cases or having a lack of validation.\(^10\)

In this study, we summarized the mutational status of EGFR mutation, subtype mutations and co-mutations in several cases, systematically identified the risk factors of EGFR mutation, and developed a prediction scoring system that is concise and readily adoptable at most institutions, aiding those who were unavailable for gene detection. Furthermore, we validated our model using prospectively collected data to examine its generalizability and reliability.

**Methods**

**Data Collection and Definitions**

This was a single-center retrospective cohort study. We performed the study in accordance with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) reporting checklist.\(^11\) The development cohort (n = 1450) was retrospectively compiled from patients with lung cancer admitted to our institution between December 2014 to October 2020. Temporal validation samples (n = 210) were prospectively collected from the same institution as the development cohort but at a later time point, between November 2020 and June 2021. The criteria for inclusion were patients with pathologically diagnosed lung cancer who underwent EGFR detection in our institution. The exclusion criteria were patients 1) whose medical records could not be obtained and 2) who were treated with any therapy before EGFR detection. We have de-identified all patient details.

We collected clinicopathological data from the electronic medical records of patients, including sex, age, smoking status, smoking index, family history of malignant tumors, history of other malignant tumors, tumor location, computerized tomography (CT) imaging manifestation, gross type, TNM and clinical stage, serum tumor markers, including carcinoembryonic antigen (CEA), alpha fetoprotein (AFP), ferritin (FERR), carbohydrate antigen (CA) 125, carbohydrate antigen (CA) 15–3, carbohydrate antigen (CA) 19–9, squamous cell carcinoma antigen (SCCA), soluble fragment of cytokeratin 19 (CYFRA21-1), carbohydrate antigen (CA) 72–4, and neuron-specific enolase (NSE), histologic subtype, differentiation grade, mucus component, and some immunohistochemical results, including thyroid transcription factor-1 (TTF-1), napsin A, P63, P40, CK-7, and Ki67. The smoking index (Brinkman index) was defined as the number of cigarettes smoked per day multiplied by the number of smoking years. Tumors were staged according to the eighth edition of the American Joint Committee on Cancer (AJCC) staging system. Positive immunohistochemical results included thyroid transcription factor-1(TTF-1), napsin A, P63, P40, CK-7 and Ki67. The smoking index (Brinkman index) was defined as the number of cigarettes smoked per day multiplied by the number of smoking years.
TNM staging classification of the International Union against Cancer and the American Joint Committee on Cancer (UICC/AJCC) for lung cancer.\textsuperscript{12} Ground-glass opacity (GGO) is defined as hazy opacity on high resolution computed tomography (HRCT), through which pulmonary vessels or bronchial structures can be visualized.\textsuperscript{12} The upper limit of each tumor marker was as follows: CEA, 5 ng/mL; AFP, 8.78 ng/mL; FERR, 130 ng/mL (male); FERR, 55 ng/mL (female); CA125, 35 U/mL; CA15-3, 31.3 U/mL; CA19-9, 37 U/mL; SCCA, 1.5 ng/mL; CYFRA 21-1, 3.3 ng/mL; CA72-4, 6.9 U/mL; and NSE 15.2 ng/mL. The above tumor markers were considered positive if their values were higher than the upper limit. Histologic subtypes were determined according to the diagnostic criteria of the 2015 WHO histological classification of lung cancer.\textsuperscript{13}

**EGFR Detection Methods**

All tumor tissue samples were fixed in 10% neutral formalin solution and embedded in paraffin. We used the QIAamp DNA formalin-fixed paraffin-embedded (FFPE) tissue kit (Qiagen NV, Venlo, Netherlands) to extract genomic DNA and RNA from FFPE tissues, according to the manufacturer’s instructions. The \textit{EGFR} analysis results were obtained and analyzed by amplification refractory mutation system-polymerase chain reaction (ARMS-PCR) or using next-generation sequencing (NGS) method.

**Development and Validation of the Scoring System**

A scoring system was developed with (a) a \( \beta \)-coefficient obtained for each significant factor from the multivariate logistical regression model (b) each of the coefficient variables rounded off to whole integers and (c) the score of each variable set as its corresponding coefficient.\textsuperscript{14} The discriminatory ability of the model was evaluated using receiver operating characteristic curve (ROC) analysis and the area under the ROC curve. Goodness-of-fit was assessed using Hosmer–Lemeshow test. The scoring system was stratified into the following three groups: low, moderate, and high risks.

The scoring system was validated in the internal and temporal validation cohorts by assessing model discrimination and calibration. The C-statistic was adopted for the quantification of discrimination, which is equal to the area under the ROC. Calibration was studied from graphical representations of the relationship between the observed outcome frequencies and the predicted probabilities (calibration curves). The internal validation involved bootstrapping techniques. Thousand bootstrap samples were drawn from the original data set with replacement. The model was temporally validated using prospectively collected data.

**Statistical Analysis**

All data were analyzed using R.4.0.5, SPSS 22.0 software (IBM Corp., Armonk, NY, USA) and Microsoft Excel (2019). Count data are expressed as the number of cases or rate (%), and the difference between groups was evaluated using the chi-square test or Fisher’s exact test. Statistical significance was set at \( P < 0.05 \). The relationships between clinicopathological characteristics and gene mutations were analyzed using univariate and multivariate logistic regression analyses. Parameters were included in the multivariate analysis only when the \( P \) value was < 0.05 in the univariate analysis.

**Results**

**Study Population**

A total of 1450 participants with lung cancer who underwent \textit{EGFR} detection from December 2014 to October 2020 were

![Figure 1. Flow diagram of patient selection for the development and temporal validation cohort.](image-url)
Figure 2. Pie charts showing the distribution of EGFR mutations in the study cohort. (A) Single mutation, (B) compound mutation.
included in our study. Furthermore, 210 samples whose data were prospectively collected between November 2020 and June 2021 were included in temporal validation cohort. A detailed flow diagram of patient selection is presented in Figure 1.

**Status of EGFR Mutation**

The identified mutations in EGFR in the 1450 patients were 723 single mutations and 51 compound mutations. The single mutations comprised 337 cases with 19 del, 322 cases with L858R, 17 cases with 20 ins, 12 cases with L861Q, 11 cases with G719X, 2 cases with S768I, and 22 cases with uncommon mutations (Figure 2A).

The compound mutations comprised 13 cases with T790M+L858R, 12 cases with T790M+19del, 5 cases with 19del+L858R, 3 cases with S768I+L858R, 3 cases with G719X+S768I, 2 cases with 19del+20ins, 2 cases with G719X+L861Q, 1 case with I706T+G719A, 1 case with E709A+G719A, 1 case with G724S+L858R, 1 case with G719S+S768I, 1 case with S720F+L858R, 1 case with 19del+L861Q, 1 case with I759M+L858R, 1 case with V819A+L858R, 1 case with T790M+E746_A750del, 1 case with L861Q+L858R, and 1 case with T790M+L858R+S768I (Figure 2B).

Furthermore, among the 1450 patients, 39 cases had 2 or more synchronous gene mutations. Among these, EGFR+TP53 was found to be the most frequent co-mutation, followed by EGFR+EML4-ALK, EGFR+BRAF, EGFR+ROS1, EGFR+KRAS, TP53+EML4-ALK, EGFR+FGFR3+TP53, KRAS+TP53, EGFR+KRAS+NRAS, EGFR+PIK3CA+FBXW7, EGFR+TP53+PTEN, EGFR+PIK3CA, EGFR+TP53+CTNNB1, EGFR+KRAS+TP53, EGFR+PTEN, EGFR+NRAS, EGFR+HRAS, KRAS+PTEN+TSC1, KRAS+STK11+TP53, PIK3CA+TP53, MET+ERBB4, ERBB2+PTEN, and ERBB2+NRAS (Figure 3).

**Relationship Between the EGFR Mutation and Clinicopathological Characteristics**

The multivariate logistic regression analysis showed that sex; smoking status; smoking index; family history of malignant tumors; history of other malignant tumors; CT imaging manifestation; gross type; levels of CEA, CA19-9, and SCCA; histologic type and subtype; differentiation grade; and levels of TTF-1 and napsin A were significantly correlated with the mutations in EGFR (Table 1).

**Development and Validation of Prediction Models**

Model 1 based on the clinical predictors demonstrated both good discrimination and calibration, as determined using the C-statistics and Hosmer–Lemeshow goodness-of-fit of 0.754 (95% CI 0.729-0.778) ($\chi^2 = 6.733$, degrees of freedom = 8, $P = 0.566$) for the development and 0.710 (95% CI 0.638-0.782) for the temporal validation cohort. The components of model 1 were sex; smoking status; smoking index; family history of malignant tumors; history of other malignant tumors; CT imaging manifestation; gross type; and levels of CEA, CA19-9, and SCCA.

Model 2 based on the clinicopathological predictors demonstrated both good discrimination and calibration, as determined using the C-statistics and Hosmer–Lemeshow goodness-of-fit of 0.812 (95% CI 0.790-0.834) ($\chi^2 = 6.418$, degrees of freedom = 8, $P = 0.989$) for the development and 0.790 (95% CI 0.730-0.851) for the temporal validation cohort. The components of model 2 were sex; smoking status; smoking index; family history of malignant tumors; history of other malignant tumors; CT imaging manifestation; gross type; levels of CEA, CA19-9, and SCCA, histologic type and subtype; differentiation grade; and levels of TTF-1 and napsin A.

Figure 4 shows the ROC curves of prediction for models 1 and 2 in the development cohort (A, C) and temporal validation cohort (B, D). The calibration of the internal validation was good, as shown in Figure 5. The x-axis is the prediction calculated from the model, and the y-axis is the actual EGFR mutation observed in the cohort. The reference line indicates the difference between
Table 1. Univariate and multivariate analysis of clinicopathological characteristics with EGFR mutational status from the development cohort.

| Characteristics                        | n     | EGFR (+) | EGFR (-) | P value | OR [95% CI] | P value |
|----------------------------------------|-------|----------|----------|---------|-------------|---------|
| **Age (y)**                            |       |          |          |         |             |         |
| Mean ± SD                              | 61.9 ± 9.8 | 61.3 ± 9.8 | 62.6 ± 9.7 | 0.015   |             | 0.378   |
| <62                                    | 657   | 376      | 281      |         |             |         |
| ≥62                                    | 793   | 398      | 395      |         |             |         |
| **Sex**                                |       |          |          | <0.001  |             | <0.001  |
| Male                                   | 876   | 358      | 518      |         | Reference   |         |
| Female                                 | 574   | 416      | 158      |         | Reference   |         |
| **Smoking status**                     |       |          |          | <0.001  |             | <0.001  |
| Current/Former                         | 771   | 233      | 538      |         |             |         |
| Never                                  | 645   | 529      | 116      |         | Reference   |         |
| **Smoking index**                      |       |          |          | <0.001  |             | 0.009   |
| Mean ± SD                              | 774.6 ± 522.3 | 664.8 ± 491.3 | 836.6 ± 529.7 |         | Reference   |         |
| ≥770                                   | 300   | 215      | 85       |         |             |         |
| <770                                   | 1116  | 677      | 439      |         | 1.646 [1.133, 2.390] |         |
| Unknown                                | 34    | 12       | 22       |         |             |         |
| **Family history of malignant tumors** |       |          |          | 0.001   | 1.701 [1.053, 2.748] | 0.030   |
| Yes                                    | 117   | 74       | 43       |         |             |         |
| No                                     | 1299  | 488      | 811      |         | Reference   |         |
| Unknown                                | 34    | 12       | 22       |         |             |         |
| **History of other malignant tumors**  |       |          |          | 0.002   | 0.249 [1.254, 4.032] | 0.007   |
| Yes                                    | 70    | 25       | 45       |         |             |         |
| No                                     | 1346  | 737      | 609      |         | Reference   |         |
| Unknown                                | 34    | 12       | 22       |         |             |         |
| Tumor location                         |       |          |          | 0.239   |             |         |
| Left                                   | 603   | 329      | 274      |         |             |         |
| Right                                  | 798   | 402      | 396      |         |             |         |
| Bilateral                              | 16    | 7        | 9        |         |             |         |
| Unknown                                | 33    | 12       | 21       |         |             |         |
| **CT imaging manifestation**           |       |          |          | <0.001  |             | 0.005   |
| Solid                                  | 1355  | 711      | 644      |         | Reference   |         |
| GGO (Pure/Mix)                         | 58    | 49       | 9        |         | 3.515 [1.463, 8.447] |         |
| Unknown                                | 37    | 14       | 23       |         |             |         |
| Gross type                             |       |          |          | <0.001  |             | 0.025   |
| Central type                           | 208   | 87       | 121      |         | Reference   |         |
| Peripheral type                        | 1205  | 673      | 532      |         | 1.535 [1.056, 2.232] |         |
| Unknown                                | 37    | 14       | 23       |         |             |         |
| **T stage**                            |       |          |          | <0.001  |             | 0.740   |
| T1                                     | 277   | 177      | 100      |         | Reference   |         |
| T2                                     | 280   | 138      | 142      |         |             |         |
| T3                                     | 143   | 65       | 78       |         | 1.057 [0.760, 1.471] |         |
| T4                                     | 237   | 114      | 123      |         |             |         |
| Unknown                                | 513   | 280      | 233      |         |             |         |
| **N stage**                            |       |          |          | 0.191   |             |         |
| N0                                     | 388   | 213      | 175      |         |             |         |
| N1                                     | 90    | 55       | 35       |         |             |         |
| N2                                     | 344   | 176      | 168      |         |             |         |
| N3                                     | 165   | 80       | 85       |         |             |         |
| Unknown                                | 463   | 250      | 213      |         |             |         |
| **M stage**                            |       |          |          | 0.328   |             |         |
| M0                                     | 537   | 276      | 261      |         |             |         |
| M1a                                    | 91    | 48       | 43       |         |             |         |
| M1b                                    | 288   | 160      | 128      |         |             |         |
| M1c                                    | 207   | 121      | 86       |         |             |         |
| Unknown                                | 327   | 169      | 158      |         |             |         |
Table 1. (continued)

| Characteristics | n | Univariate analysis | Multivariate analysis* |
|-----------------|---|---------------------|------------------------|
|                 |   | EGFR (+) | EGFR (-) | P value | OR [95% CI] | P value |
| Clinical stage  |   |          |          |         |            |         |
| I A             | 155 | 100    | 55      | 0.019   |            |         |
| I B             | 52  | 24     | 28      | 0.142   |            |         |
| II A            | 25  | 8      | 17      | 0.094   |            |         |
| II B            | 80  | 39     | 41      | 0.436   |            |         |
| III A           | 124 | 65     | 59      | 0.094   |            |         |
| III B           | 81  | 33     | 48      | 0.094   |            |         |
| III C           | 22  | 7      | 15      | 0.436   |            |         |
| IV A            | 390 | 215    | 175     | 0.436   |            |         |
| IV B            | 207 | 121    | 86      | 0.436   |            |         |
| I-III           | 814 | 423    | 391     | 0.094   |            |         |
| IV              | 597 | 336    | 261     | 0.094   |            |         |
| Unknown         | 39  | 25     | 14      | 0.436   |            |         |
| CEA             |     | 0.011  |         | 0.002   | 1.627 [1.195, 2.217] |         |
| Positive        | 578 | 332    | 246     | 0.011   |            |         |
| Negative        | 659 | 331    | 328     | 0.011   |            |         |
| Unknown         | 213 | 111    | 102     | 0.011   |            |         |
| AFP             |     | 0.724  |         |         |            |         |
| Positive        | 23  | 11     | 12      | 0.724   |            |         |
| Negative        | 803 | 414    | 389     | 0.724   |            |         |
| Unknown         | 624 | 349    | 275     | 0.724   |            |         |
| FERR            |     | 0.387  |         |         |            |         |
| Positive        | 68  | 29     | 39      | 0.387   |            |         |
| Negative        | 147 | 72     | 75      | 0.387   |            |         |
| Unknown         | 1235 | 673 | 562     | 0.387   |            |         |
| CA125           |     | 0.031  |         | 0.624   | 1.085 [0.783, 1.503] |         |
| Positive        | 455 | 226    | 229     | 0.031   |            |         |
| Negative        | 729 | 409    | 320     | 0.031   |            |         |
| Unknown         | 266 | 139    | 127     | 0.031   |            |         |
| CA15-3          |     | 0.074  |         |         |            |         |
| Positive        | 86  | 57     | 29      | 0.074   |            |         |
| Negative        | 382 | 213    | 169     | 0.074   |            |         |
| Unknown         | 982 | 504    | 478     | 0.074   |            |         |
| CA19-9          |     | 0.030  |         | 0.029   | Reference  |         |
| Positive        | 152 | 66     | 86      | 0.030   |            |         |
| Negative        | 687 | 365    | 322     | 0.030   |            |         |
| Unknown         | 611 | 343    | 268     | 0.030   |            |         |
| SCCA            |     | <0.001 |         | 0.006   | Reference  |         |
| Positive        | 78  | 19     | 59      | <0.001  |            |         |
| Negative        | 273 | 157    | 116     | <0.001  |            |         |
| Unknown         | 1099 | 598 | 501     | <0.001  |            |         |
| CYFRA 21-1      |     | 0.004  |         | 0.881   | Reference  |         |
| Positive        | 206 | 92     | 114     | 0.004   |            |         |
| Negative        | 130 | 79     | 51      | 0.004   |            |         |
| Unknown         | 1114 | 603 | 511     | 0.004   |            |         |
| CA72-4          |     | 0.163  |         |         |            |         |
| Positive        | 45  | 17     | 28      | 0.163   |            |         |
| Negative        | 145 | 72     | 73      | 0.163   |            |         |
| Unknown         | 1260 | 685 | 575     | 0.163   |            |         |
| NSE             |     | 0.252  |         |         |            |         |
| Positive        | 237 | 121    | 116     | 0.252   |            |         |
| Negative        | 869 | 480    | 389     | 0.252   |            |         |
| Unknown         | 344 | 173    | 171     | 0.252   |            |         |
| Adenocarcinoma component |     | <0.001 |         | 0.001   | Reference  |         |
| With            | 1313 | 758 | 555     | <0.001  |            |         |
| Without         | 117  | 15     | 102     | <0.001  |            |         |

(continued)
| Characteristics                          | n   | EGFR (+) | EGFR (-) | P value | OR [95% CI] | P value |
|-----------------------------------------|-----|----------|----------|---------|-------------|---------|
| NOS#                                    | 9   |          |          | 9       |             |         |
| Unknown                                 | 11  |          |          |         |             |         |
| Predominant component in adenocarcinoma |     |          |          | <0.001  |             |         |
| Micropapillary/Solid                    | 78  | 25       | 53       |         |             |         |
| Acinar/Papillary                        | 211 | 141      | 70       |         | 2.962 [1.587, 5.528] | 0.001  |
| AIS/Lepidic                             | 73  | 53       | 20       |         | 2.983 [1.316, 6.766] | 0.009  |
| Minimally invasive                      | 2   | 0        | 2        |         |             |         |
| Unknown                                 | 909 | 521      | 388      |         |             |         |
| Adenocarcinoma                          | 1273| 740      | 533      | 0.098   |             |         |
| Adenosquamous                           | 40  | 18       | 22       |         |             |         |
| Differentiation grade                   |     | <0.001   |          |         |             |         |
| Poor                                    | 271 | 86       | 185      |         | Reference   |         |
| Moderate                                | 241 | 158      | 83       |         | 2.566 [1.657, 3.973] | <0.001 |
| Well                                    | 70  | 50       | 20       |         | 3.131 [1.590, 6.165] | 0.001  |
| Unknown                                 | 868 | 480      | 388      |         |             |         |
| Mucus component                         |     | 0.005    | 0.073    |         |             |         |
| Yes                                     | 36  | 11       | 25       |         | 0.457 [0.194, 1.075] |         |
| No                                      | 1414| 763      | 651      |         | Reference   |         |
| TTF-1                                   |     | <0.001   |          |         |             |         |
| Positive                                | 866 | 483      | 383      |         | 2.853 [1.480, 5.500] | 0.002  |
| Negative                                | 138 | 17       | 121      |         | Reference   |         |
| Unknown                                 | 446 | 274      | 172      |         |             |         |
| Napsin A                                |     | <0.001   | <0.001   |         |             |         |
| Positive                                | 648 | 374      | 274      |         | 3.003 [1.775, 5.083] |         |
| Negative                                | 193 | 29       | 164      |         | Reference   |         |
| Unknown                                 | 609 | 371      | 238      |         |             |         |
| P63                                     |     | 0.974    |          |         |             |         |
| Positive                                | 210 | 61       | 149      |         |             |         |
| Negative                                | 294 | 85       | 209      |         |             |         |
| Unknown                                 | 946 | 628      | 318      |         |             |         |
| P40                                     |     | 0.009    | 0.441    |         |             |         |
| Positive                                | 67  | 21       | 46       |         | 1.340 [0.637, 2.822] |         |
| Negative                                | 360 | 175      | 185      |         | Reference   |         |
| Unknown                                 | 1023| 578      | 445      |         |             |         |
| CK-7                                    |     | <0.001   | 0.280    |         |             |         |
| Positive                                | 687 | 343      | 344      |         | 0.549 [0.185,1.629] |         |
| Negative                                | 42  | 8        | 34       |         | Reference   |         |
| Unknown                                 | 721 | 423      | 298      |         |             |         |
| Ki67 (%)                                |     | <0.001   | 0.052    |         |             |         |
| Mean ± SD                               | 36.4 ± 21.6 | 32.9 ± 20.9 | 40.0 ± 21.8 |         | Reference   |         |
| <36                                     | 461 | 264      | 197      |         |             |         |
| ≥36                                     | 453 | 205      | 248      |         | 1.396 [0.996, 1.956] |         |
| Unknown                                 | 536 | 305      | 231      |         |             |         |
| Specimen type                           |     | 0.140    |          |         |             |         |
| Biopsy                                  | 811 | 419      | 392      |         |             |         |
| Surgical resection                      | 639 | 355      | 284      |         |             |         |
| Technology                              |     | 0.430    |          |         |             |         |
| ARMS                                    | 1167| 617      | 550      |         |             |         |
| NGS                                     | 283 | 157      | 126      |         |             |         |

Abbreviations: AIS: adenocarcinoma in situ, AFP: alpha fetoprotein, ARMS: amplification refractory mutation system, CA: carbohydrate antigen, CEA: carcinoembryonic antigen, CI: confidence interval, CK: cytokeratin, CT: computerized tomography, EGFR: epidermal growth factor receptor, FERR: ferritin, GGO: ground-glass opacity, NGS: next-generation sequencing, NSE: neuron-specific enolase, OR: odds ratio, SCCA: squamous cell carcinoma antigen, SD: standard deviation, CYFRA21-1: soluble fragment of cytokeratin 19, TTF-1: thyroid transcription factor-1.

*Items were included in the multivariate analysis only when the P value is <0.05 in univariate analysis.

#NOS, not otherwise specified indicates pathologically confirmed NSCLC, but the pathologic type was not clearly identified.
the predicted probabilities. The calibration plots showed a small deviation from ideal predictions.

As shown in Table 2, there was no significant difference in the prediction accuracy between the models (P = 0.455).

**Construction of a Clinical Scoring System**

We developed a concise scoring system enabling the differentiation of *EGFR* mutations based on the β coefficients obtained for the independent clinical predictors from the multivariate logistical regression model (Table 3). The predictive equation was logistic \( (p) = -4.208 + (0.802 \times \text{"female"}) + (0.625 \times \text{"never smoking"}) + (0.591 \times \text{"smoking index<770"}) + (0.592 \times \text{"with family history of malignant tumors"}) + (0.973 \times \text{"without history of other malignant tumors"}) + (1.314 \times \text{"ground-glass opacity"}) + (0.555 \times \text{"peripheral type"}) + (0.659 \times \text{"positive CEA"}) + (0.567 \times \text{"negative CA19-9"}) + (1.180 \times \text{"negative SCCA"}) \). Model scores ranged from 0 to 13.

**Risk Group Stratification and Probability of EGFR Mutation**

Patients were stratified into risk groups according to their scores: low-risk (score <4), moderate-risk (score 4-8), and high-risk (score >8) groups. In the development cohort, compared

---

**Figure 4.** Receiver operating characteristic curve of models 1 and 2 in the development cohort (A,C) and temporal validation cohort (B,D).
Figure 5. Calibration plot comparing the actual and predicted probabilities of the EGFR mutation. (A) Model 1; (B) model 2.
with that in the low-risk group, the occurrence of \( \text{EGFR} \) mutation was increased in the moderate-risk (OR = 5.22, 95% CI 3.36-8.11, \( P < 0.001 \)) and high-risk (OR = 19.67, 95% CI 11.22-34.50, \( P < 0.001 \)) groups. In the temporal validation cohort, compared with that in the low-risk group, the occurrence of \( \text{EGFR} \) mutation was increased in the moderate-risk (OR = 1.91, 95% CI 0.75-4.85, \( P = 0.172 \)) and high-risk (OR = 8.08, 95% CI 2.80-23.33, \( P < 0.001 \)) groups (Table 4).

Figure 6 presents the probability of \( \text{EGFR} \) mutation. In the development cohort, among patients with low, moderate, and high risks of \( \text{EGFR} \) mutation, the actual mutation frequency was 18.2%, 53.7%, and 81.4%, respectively. In the temporal validation cohort, among patients with low, moderate, and high risks of \( \text{EGFR} \) mutation, the actual mutation frequency was 28.0%, 42.6%, and 68.2%, respectively.

### Discussion

There were two main parts to our study. In the first part, we described the \( \text{EGFR} \) genetic profile of lung cancer. We observed that among 1450 lung cancer cases, the most common mutation was identified in \( \text{EGFR} \). More specifically, mutations in \( \text{EGFR} \) were detected in more than half of the patients (774/1450, 53.4%). A high frequency of mutations in \( \text{EGFR} \) in Chinese patients with lung cancer highlights the significance of exploring the mutational status of \( \text{EGFR} \).3,5 In our study, the most common mutational subtype in \( \text{EGFR} \) was 19 del (46.6%), followed by L858R (44.5%). However, several studies have reported that the frequency of L858R was higher than that of 19 del,15 which might have been due to the number of samples, geographical differences, various tumor stages, different specimen types, and detection technologies. In addition, T790M was confirmed as the most important

### Table 2. Comparison of prediction accuracy between two models.

| Model | Successful prediction | Failure prediction | P value |
|-------|-----------------------|--------------------|---------|
| Model 1 | 144 | 66 | 0.455 |
| Model 2 | 151 | 59 |         |

### Table 3. Multivariate analysis for the independent clinical predictors in EGFR mutation and corresponding points.

| Categories | \( \beta \) | S.E. | Wald | P value | OR [95% CI] | Points |
|------------|-------------|------|------|---------|-------------|--------|
| Intercept  | -4.208      | 0.541| 60.496| <0.001  | Reference   | 0      |
| Sex        |             |      |       |         |             | 1      |
| Male       |             |      |       |         |             |        |
| Female     | 0.802       | 0.149| 29.150| <0.001  | 2.230 [1.667, 2.984] | 1 |
| Smoking status |              |      |       |         | Reference   | 0      |
| Current/Former |       |      |       |         |             | 1      |
| Never      | 0.625       | 0.162| 14.807| <0.001  | 1.869 [1.359, 2.570] | 1 |
| Smoking index |             |      |       |         | Reference   | 0      |
| \( \geq 770 \) |             |      |       |         |             | 1      |
| \(< 770 \)  | 0.591       | 0.177| 11.155| 0.001   | 1.806 [1.277, 2.555] | 1 |
| Family history of malignant tumors | | | | | | |
| No         |             |      |       |         | Reference   | 0      |
| Yes        | 0.592       | 0.226| 6.877 | 0.009   | 1.808 [1.161, 2.815] | 1 |
| History of other malignant tumors | | | | | | |
| No         |             |      |       |         | Reference   | 0      |
| Yes        | 0.973       | 0.279| 12.213| <0.001  | 2.647 [1.533, 4.570] | 2 |
| CT imaging manifestation | | | | | | |
| Solid      |             |      |       |         | Reference   | 0      |
| GGO (Pure/Mix) | 1.314  | 0.392| 11.687| 0.001   | 3.822 [1.772, 8.243] | 2 |
| Gross type |             |      |       |         | Reference   | 0      |
| Central type |             |      |       |         |             | 1      |
| Peripheral type | 0.555 | 0.173| 10.282| 0.001   | 1.743 [1.241, 2.447] | 1 |
| CEA        |             |      |       |         | Reference   | 0      |
| Negative   |             |      |       |         |             | 1      |
| Positive   | 0.659       | 0.144| 21.048| <0.001  | 1.932 [1.458, 2.561] | 1 |
| CA19-9     |             |      |       |         | Reference   | 0      |
| Positive   | 0.567       | 0.212| 7.149 | 0.008   | 1.763 [1.163, 2.671] | 1 |
| SCCA       |             |      |       |         | Reference   | 0      |
| Positive   | 1.180       | 0.316| 13.903| <0.001  | 3.253 [1.750, 6.049] | 2 |
| Total      |             |      |       |         | Reference   | 13     |

Abbreviations: CA: carbohydrate antigen, CEA: carcinoembryonic antigen, CI: confidence interval, CT: computerized tomography, EGFR: epidermal growth factor receptor, GGO: ground-glass opacity, OR: odds ratio, SCCA: squamous cell carcinoma antigen.
TKI-resistant mutation, accounting for more than half of the observed secondary resistance. We identified one instance with single T790M mutation, whereas all other instances were cases of coexistence with sensitive mutation sites. Furthermore, T790M was found to incline to co-mutate with L858R (25.5%) compared with 19 del (23.5%), suggesting that point mutations are more likely to have compound mutations. In our cohort, all patients with a T790M secondary mutation were treated with osimertinib, an irreversible third generation TKI drug.17

EGFR/TP53 co-mutation was found to be the most frequent in two or more synchronous gene mutations. TP53, as a tumor suppressor gene, is frequently identified in lung adenocarcinoma and is associated with a poor prognosis of patients with non-small cell lung cancer (NSCLC).18 Mutations in TP53 have been reported to result in DNA repair and excessive apoptosis, leading to cancer.19 A retrospective study of 1017 samples found that two of the most common mutations in the EGFR-mutant lung cancer were PIK3CA and MET,14 but this study did not analyze the tumor suppressor genes. Zhao et al. reported that patients with co-mutations in EGFR and TP53 accounted for 22.4% of all patients with lung adenocarcinoma and had higher tumor mutational burden and worse recurrence-free survival. In addition, the authors provided insights into the prognostic value of co-mutation of EGFR/TP53 for patients with lung adenocarcinoma.20 Thus, patients with lung cancer harboring EGFR and co-mutational tumor suppressor genes should be regarded as a unique subgroup, and more informed and genomically empowered molecular diagnosis and monitoring, and dynamically applied rational polytherapy strategies should be employed to address the clonal and subclonal co-alterations driving disease progression and drug resistance in order to better control lung cancer.

In the second part, we established two prediction models based on the clinical and clinicopathological features and found there was no significant difference between them (P = 0.455). Thus, for patients who are not available for gene detection pathologically, we developed a practical predictive scoring system that could initially identify EGFR mutation, based on the independent clinical predictors, including sex, smoking status, smoking index, family history of malignant tumors, history of other malignant tumors, CT imaging manifestation, gross type, CEA, CA19-9, and SCCA. The discrimination of our scoring system in distinguishing patients with wild-type or EGFR-mutated tumors was good. The system uses coefficients to convert the features into quantitative data, demonstrating good discrimination with C-statistics of 0.754 (95% CI 0.729-0.778) for the development and 0.710 (95% CI 0.638-0.782) for the temporal validation cohorts.

The system demonstrated that mutations in EGFR were significantly associated with female sex and non-smokers or smokers with lower smoking index (<770), but not stage. A study of 884 patients reported that mutations in EGFR are common in Chinese patients with early-stage non-small cell lung cancer (NSCLC; I–II);5 whereas another study reported a contradictory conclusion.2 Most surgical specimens reflect cases of early lung cancer, whereas most biopsy specimens reflect cases of advanced lung cancer. Therefore, inclusion of only a single specimen type is likely to be the underlying cause of the discordance in the results between these studies. Moreover, our study showed that biopsy specimens can replace surgical specimens for the detection of the mutational status of EGFR (P > 0.05), consistent with the findings of a previous study.16 In our study, we included the smoking index (Brinkman index), which can reflect the degree of influence of smoking.9 Approximately 83% of women with lung cancer do not have a smoking habit in China. Sex may influence the correlation between smoking status and EGFR mutation. Tomita et al.20 compared the correlation between the Brinkman index and EGFR mutation status of 90 Japanese men and found that the Brinkman index of the EGFR mutation

### Table 4. Scoring system table.

| Risk score quantities | n   | EGFR (+), n (%) | OR [95% CI] | P  |
|-----------------------|-----|----------------|-------------|----|
| Development Cohort    |     |                |             |    |
| Range (0, 12)         |     |                |             |    |
| Score ≤4              | 143 | 26 (18.2)      | Reference   |    |
| 4 ≤ Score ≤8          | 1108| 595 (53.7)     | 5.22 [3.36, 8.11] | <0.001 |
| Score >8              | 188 | 153 (81.4)     | 19.67 [11.22, 34.50] | <0.001 |
| Temporal Validation   |     |                |             |    |
| Cohort                |     |                |             |    |
| Range (1, 13)         |     |                |             |    |
| Score ≤4              | 25  | 7 (28.0)       | Reference   |    |
| 4 ≤ Score ≤8          | 141 | 60 (42.6)      | 1.91 [0.75, 4.85] | 0.172  |
| Score >8              | 44  | 30 (68.2)      | 8.08 [2.80, 23.33] | <0.001 |

Abbreviations: CI: confidence interval, EGFR: epidermal growth factor receptor, OR: odds ratio.

![Figure 6](image-url)  
**Figure 6.** Histogram for the risk of EGFR mutation. According to the scoring system, a histogram of the proportions of the low-, medium-, and high-risk populations were drawn.
group was lower than that of the wild-type group, but with no significant differences ($P = 0.8357$). A recent Chinese study$^9$ analyzed the Brinkman index in male subjects and found that the degree of smoking between the $EGFR$ group and the control group was still significantly different ($P = 0.020$). After excluding the influence of sex, our result was in agreement with the latter. A meta-analysis also showed that TKI therapy should be expanded to former smokers with less than 15 pack-years or those who did not smoke for more than 25 years.$^{21}$ In addition, a family history of malignant tumors and history of other malignant tumors were found to be significantly related to mutations in $EGFR$. To the best of our knowledge, there have been no reports on these factors.

Our system showed that a higher proportion of GGO was linked with mutations in $EGFR$, consistent with most previous study findings.$^{8,9,22}$ However, some previous studies have reported that either no significant correlation was observed between mutations in $EGFR$ and GGO$^{23}$ or mutations in $EGFR$ were more common in lesions with >50% solid component.$^{24,25}$ The difference could be a result of insufficient recruitment of early-stage patients, or because different definitions were used for the quantification and classification of GGO.$^8$

After excluding the influence of other malignant tumors, positive CEA, negative SCCA, and negative CA19-9 in serum were demonstrated to be predictors of mutations in $EGFR$. Niu et al. indicated that compared with the wild type, patients with $EGFR$ mutations were more likely to have a high CEA level.$^{26}$ Moreover, high CEA levels may represent higher tumor burden; therefore, it has been recommended to successfully evaluate the efficacy and prognosis of treatment with $EGFR$-TKIs.$^{27}$ TKIs are known to disrupt abnormal downstream signaling pathways induced by mutations in $EGFR$ to inhibit tumor proliferation, whereas overexpression of CEA has been reported to accelerate tumor development, suggesting that CEA might be a "cofactor" of mutations in $EGFR$. However, the opposite conclusion, that is, increased serum tumor markers are independent of driver genes, has also been drawn.$^{28}$ Therefore, further research needs to be performed in this regard. Besides, SCCA is known to be the preferred marker of squamous cell carcinoma, whereas mutations in $EGFR$ are common in adenocarcinoma. Thus, detecting mutations in $EGFR$ in patients with expressing of SCCA would be rare, consistent with the univariate analysis results of Wen et al.$^{29}$ However, the inconsistency with their multivariate analysis results may be related to sample size, inclusion criteria, and population differences. We also demonstrated the value of CA 19-9 for the prediction of $EGFR$ mutations in patients with lung cancer. To date, no relevant research has been conducted. In addition, a study showed that cytological CYFRA21-1 correlated with the mutational status in $EGFR$, but not after its release into the serum.$^{30}$ Therefore, serum tumor markers, such as CEA, SCCA, and CA19-9, the detection of which is both convenient and easy in China, could be widely used in hospitals lacking gene detection technology.

The proposed scoring system is based on clinical characteristics with external validation to show its generalizability and reliability. According to our scoring system, low-, moderate-, and high-risk grades may distinguish the $EGFR$ mutation type and wild type. Although a few studies have evaluated the risk factors of $EGFR$ mutation or propose a scoring system, the parameters were incomprehensive and the efficiency of these scoring systems was not externally validated.$^{2,10,15}$ To the best of our knowledge, the current study proposes the first scoring system using various clinical parameters to predict the $EGFR$ mutation status in a relatively large development cohort and validate its clinical value.

Regarding pathological features, we found that patients with well-differentiated adenocarcinoma and peripheral tumors were more likely to have $EGFR$ mutations. The $EGFR$ mutations were more common in patients with AIS or lepidic predominant adenocarcinoma. It has been reported that the CT findings of GGO tend to correspond to lepidic pattern observed pathologically, although this correlation is not absolute.$^{23}$ The prognosis was best for AIS and lepidic predominant adenocarcinoma. The OS rate of patients with micropapillary or solid predominant adenocarcinoma was significantly lower than that of patients with minimally invasive, lepidic, acinar, or papillary adenocarcinoma. Patients with acinar and papillary adenocarcinoma exhibit a similar cumulative 1-year overall survival (OS) rate.$^{32}$ Thus, we classified different subtypes of adenocarcinoma based on the corresponding prognosis when constructing the model. The results revealed that the $EGFR$ mutation rate was the highest in patients with AIS/lepidic predominant adenocarcinoma, followed by patients with papillary/acinar predominant adenocarcinoma and micropapillary/solid predominant adenocarcinoma. Previous study findings indicated that $EGFR$-mutated tumors are significantly associated with lepidic or acinar predominant subtype.$^{32-34}$ Thus, when performing subsolid lung nodule biopsy for $EGFR$ mutation detection, it is recommended that interventional radiologists should target the GGO components as much as possible to increase the detection rate of $EGFR$ mutations.

Additionally, TTF-1 and napsin A have been demonstrated to be significant predictive factors for $EGFR$ mutation. TTF-1 is a tissue-specific transcription factor that helps to maintain the functions of terminal respiratory unit cells.$^{35}$ Among immunohistochemical markers, TTF-1 is considered the gold standard for primary lung adenocarcinoma.$^{36}$ Napsin A is an apasrtic protease expressed in the lung and kidney.$^{37}$ that is capable of cleaving the preform of surfactant protein B expressed in type II pneumocytes.$^{38}$ It has been reported that napsin A is a more sensitive marker than TTF-1 for pulmonary ADCs, which may complement immunohistochemical analyses using TTF-1.$^{39,40}$ Hence, the immunohistochemical results might assist in guiding targeted therapy.$^{35,41}$

This study had some limitations. First, a retrospective single-center study could only reflect the mutational rate in patients in local areas in Asia. Second, we did not analyze the clinical value of radiomics signature or establish $EGFR$ subtype prediction model. Third, owing to a small amount of $EGFR$ compound mutations and co-mutations, we only described statistical association. Finally, NGS technology is expensive,
and most patients in our study could not afford it. Our research data on rare gene mutations was limited.

**Conclusions**

In conclusion, we determined the EGFR genetic profile of lung cancer in Asians and systematically analyzed the relevant clinicopathological factors. We developed and validated a concise and non-invasive scoring system based on the clinical predictors to initially predict the EGFR mutation status in those who are not available for gene detection.

**Acknowledgments**

We would like to thank Editage (www.editage.cn) for English language editing.

**Declaration of Conflicting Interests**

The Authors declare that there is no conflict of interest.

**Ethics Approval and Consent to Participate**

The study protocol was approved by the Medical Ethics Committee of Zhongnan Hospital of Wuhan University, China (Ethical number: 2021057). There’s no need for informed consent in our study since the unidentified data were free from medical ethics review.

**Funding**

This work was supported by Clinical Research and Development Project of Zhongnan Hospital of Wuhan University (NO. lcyf202104).

**Authors’ Contributions**

WT A, SF T, and MY L conceived and designed the study. WT A, W F, FY Z, BC W, S W, and T G collected the data. WT A and FY Z analyzed the data and provided recommendations regarding statistical analyses. WT A and W F conducted literature search. WT A, FY Z, and BC W generated the tables and figures. WT A wrote the manuscript. MY L and SF T helped to critically review and comprehensively revise the manuscript. MY L supervised the research. All authors contributed to the research and approved the final version of this article.

**ORCID iD**

Meiyan Liao 📘 https://orcid.org/0000-0002-9515-6635

**References**

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020 Jan;70(1):7-30.
2. Gaur P, Bhattacharya S, Kant S, et al. EGFR mutation detection and its association with clinicopathological characters of lung cancer patients. *World J Oncol*. 2018 Nov;9(5-6):151-155.
3. Wang S, Ma P, Ma G, et al. Value of serum tumor markers for predicting EGFR mutations and positive ALK expression in 1089 Chinese non-small-cell lung cancer patients: a retrospective analysis. *Eur J Cancer*. 2020 Jan;124:1-14.
4. Liu J, Liu Y. Molecular diagnostic characteristics based on the next generation sequencing in lung cancer and its relationship with the expression of PD-L1. *Pathol Res Pract*. 2020 Feb;216(2):152797.
5. Li D, Ding L, Ran W, et al. Status of 10 targeted genes of non-small cell lung cancer in eastern China: a study of 884 patients based on NGS in a single institution. *Thorac Cancer*. 2020 Sep;11(9):2580-2589.
6. Arriola E, García Gómez R, Diz P, et al. Clinical management and outcome of patients with advanced NSCLC carrying EGFR mutations in Spain. *BMC Cancer*. 2018 Jan 30;18(1):106.
7. Taylor C, Chacko S, Davey M, et al. Peptide-affinity precipitation of extracellular vesicles and cell-free DNA improves sequencing performance for the detection of pathogenic mutations in lung cancer patient plasma. *Int J Mol Sci*. 2020 Nov 29;21(23):9083.
8. Choi Y, Kim KJ, Jeong BH, et al. Clinicoradiopathological features and prognosis according to genomic alterations in patients with resected lung adenocarcinoma. *J Thorac Dis*. 2020 Oct;12(10):5357-5368.
9. Lv J, Zhang H, Ma J, et al. Comparison of CT radiogenomic and clinical characteristics between EGFR and KRAS mutations in lung adenocarcinomas. *Clin Radiol*. 2018 Jun;73(6):590.e1-590.e8.
10. Cao Y, Xu H. A new predictive scoring system based on clinical data and computed tomography features for diagnosing EGFR-mutated lung adenocarcinoma. *Curr Oncol*. 2018 Apr;25(2):e132-e138.
11. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br Med J*. 2015 Jan 7;350:g7594.
12. Rami-Porta R, Asamura H, Travis WD, et al. Lung cancer - major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin*. 2017 Mar;67(2):138-155.
13. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol*. 2015 Sep;10(9):1243-1260.
14. Sullivan LM, Massaro JM, D’Agostino RB, et al. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med*. 2004 May 30;23(10):1631-1660.
15. Ning H, Liu M, Wang L, et al. Clinicopathological features of Chinese lung cancer patients with epithelial growth factor receptor mutation. *J Thorac Dis*. 2017 Mar;9(3):796-801.
16. Hayashi T, Kohnsaka S, Takamochi K, et al. Clinicopathological characteristics of lung adenocarcinoma with compound EGFR mutations. *Hum Pathol*. 2020 Sep;103:42-51.
17. Zheng HY, Wang HB, Shen FJ, et al. EGFR gene mutation and methodological evaluation in 399 patients with non-small cell lung cancer. *Curr Med Sci*. 2020 Feb;40(1):78-84.
18. La Fleur L, Falk-Sörgqvist E, Smeds P, et al. Mutation patterns in a population-based non-small cell lung cancer cohort and prognostic impact of concomitant mutations in KRAS and TP53 or STK11. *Lung Cancer*. 2019 Apr;130:50-58.
19. Zhao Y, Pan Y, Cheng C, et al. EGFR-mutant lung adenocarcinoma harboring co-mutational tumor suppressor genes predicts poor prognosis. *J Cancer Res Clin Oncol*. 2020 Jul;146(7):1781-1789.
20. Tomita M, Ayabe T, Chosa E, et al. Epidermal growth factor receptor mutations in Japanese men with lung adenocarcinomas. *Asian Pac J Cancer Prev*. 2014;15(24):10627-10630.

21. Wang S, Wang Z. EGFR Mutations in patients with non-small cell lung cancer from mainland China and their relationships with clinicopathological features: a meta-analysis. *Int J Clin Exp Med*. 2014 Aug 15;7(8):1967-1978.

22. Sabri A, Batool M, Xu Z, et al. Predicting EGFR mutation status in lung cancer: proposal for a scoring model using imaging and demographic characteristics. *Eur Radiol*. 2016;26(11):4141-4147.

23. Qin X, Gu X, Lu Y, et al. EGFR-TKI-sensitive mutations in lung carcinomas: are they related to clinical features and CT findings? *Cancer Manag Res*. 2018 Oct 1;10:4019-4027.

24. Glynn C, Zakowski MF, Ginsberg MS. Are there imaging characteristics associated with epidermal growth factor receptor and KRAS mutations in patients with adenocarcinoma of the lung with bronchioalveolar features? *J Thorac Oncol*. 2010 Mar;5(3):344-348.

25. Wang T, Zhang T, Han X, et al. Impact of the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification of stage IA adenocarcinoma of the lung: Correlation between computed tomography images and EGFR and KRAS gene mutations. *Exp Ther Med*. 2015 Jun;9(6):2095-2103.

26. Niu L, Dang C, Li L, et al. Next-generation sequencing-based identification of EGFR and NOTCH2 complementary mutations in non-small cell lung cancer. *Oncol Lett*. 2021 Aug;22(2):594.

27. Ohtaki Y, Shimizu K, Suzuki H, et al. Japanese Association for Chest Surgery. Salvage surgery for non-small cell lung cancer after tyrosine kinase inhibitor treatment. *Lung Cancer*. 2021 Mar;153:108-116.

28. Noonan SA, Patil T, Gao D, et al. Baseline and on-treatment characteristics of serum tumor markers in stage IV oncogene-addicted adenocarcinoma of the lung. *J Thorac Oncol*. 2018 Jan;13(1):134-138.

29. Wen L, Wang S, Xu W, et al. Value of serum tumor markers for predicting EGFR mutations in non-small cell lung cancer patients. *Ann Diagn Pathol*. 2020 Dec;49:151633.

30. Cho A, Hur J, Moon YW, et al. Correlation between EGFR gene mutation, cytologic tumor markers, 18F-FDG uptake in non-small cell lung cancer. *BMC Cancer*. 2016 Mar 16;16:224.

31. Travis WD, Asamura H, Bankier AA, et al. International association for the study of lung cancer staging and prognostic factors committee and advisory board members. The IASLC lung cancer staging project: proposals for coding T categories for subsolid nodules and assessment of tumor size in part-solid tumors in the forthcoming eighth edition of the TNM classification of lung cancer. *J Thorac Oncol*. 2016 Aug;11(8):1204-1223.

32. Deng H, Liu J, Duan X, et al. The relationship between EGFR mutation status and clinic-pathologic features in pulmonary adenocarcinoma. *Pathol Res Pract*. 2018 Mar;214(3):450-454.

33. Wang H, Schabath MB, Liu Y, et al. Clinical and CT characteristics of surgically resected lung adenocarcinomas harboring ALK rearrangements or EGFR mutations. *Eur J Radiol*. 2016 Nov;85(11):1934-1940.

34. Villa C, Cagle PT, Johnson M, et al. Correlation of EGFR mutation status with predominant histologic subtype of adenocarcinoma according to the new lung adenocarcinoma classification of the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society. *Arch Pathol Lab Med*. 2014 Oct;138(10):1353-1357.

35. Kim HS, Kim JH, Han B, et al. Correlation of thyroid transcription factor-1 expression with EGFR mutations in non-small-cell lung cancer: a meta-analysis. *Medicina (Kaunas)*. 2019 Feb 7;55(2):41.

36. Rekhtman N, Ang DC, Sima CS, et al. Immunohistochemical algorithm for differentiation of lung adenocarcinoma and squamous cell carcinoma based on large series of whole-tissue sections with validation in small specimens. *Mod Pathol*. 2011 Oct;24(10):1348-1359.

37. Chuman Y, Bergman A, Ueno T, et al. Napsin A, a member of the aspartic protease family, is abundantly expressed in normal lung and kidney tissue and is expressed in lung adenocarcinomas. *FEBS Lett*. 1999 Nov 26;462(1-2):129-134.

38. Ueno T, Linder S, Na CL, et al. Processing of pulmonary surfactant protein B by napsin and cathepsin H. *J Biol Chem*. 2004 Apr 16;279(16):16178-16184.

39. Bishop JA, Sharma R, Illei PB. Napsin A and thyroid transcription factor-1 expression in carcinomas of the lung, breast, pancreas, colon, kidney, thyroid, and malignant mesothelioma. *Hum Pathol*. 2010 Jan;41(1):20-25.

40. Turner BM, Cagle PT, Sainz IM, et al. Napsin A, a new marker for lung adenocarcinoma, is complementary and more sensitive and specific than thyroid transcription factor 1 in the differential diagnosis of primary pulmonary carcinoma: evaluation of 1674 cases by tissue microarray. *Arch Pathol Lab Med*. 2012 Feb;136(2):163-171.

41. Yao G, Zhou Y, Gu Y, et al. Value of combining PET/CT and clinicopathological features in predicting EGFR mutation in Lung Adenocarcinoma with Bone Metastasis. *J Cancer*. 2020 Jul 11;11(18):5511-5517.