Original Article

CT-based decision tree model for predicting EGFR mutation status in synchronous multiple primary lung cancers

Yingwei Luo[1#], Shuangjiang Li[2#], Huiyun Ma[1], Wenbiao Zhang[1], Baocong Liu[1], Chuanmiao Xie[1], Qiong Li[1]

[1]Department of Radiology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, China; [2]Department of Endoscopy, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, China

Contributions: (I) Conception and design: Q Li, C Xie; (II) Administrative support: C Xie, Q Li; (III) Provision of study materials or patients: Q Li, C Xie; (IV) Collection and assembly of data: Y Luo, S Li, H Ma, W Zhang, B Liu; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Qiong Li; Chuanmiao Xie. Department of Radiology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, China. Email: liqiong@sysucc.org.cn; xchuanm@sysucc.org.cn.

Background: The current study aimed to construct a computed tomography (CT)-based decision tree algorithm (DTA) model to predict the epidermal growth factor receptor (EGFR) mutation status in synchronous multiple primary lung cancers (SMPLCs).

Methods: The demographic and CT findings of 85 patients with molecular profiling for surgically resected SMPLCs were reviewed retrospectively. Least absolute shrinkage and selection operator (LASSO) regression was used to select the potential predictors of EGFR mutation, and a CT-DTA model was developed. Multivariate logistic regression analysis and receiver operating characteristic (ROC) curve analysis were performed to assess the performance of this CT-DTA model.

Results: The CT-DTA model was applied to predict the EGFR mutant that had ten binary split, of which eight parameters to accurately categorize the lesions as follows: the presence of bubble-like vacuole sign (19.4% importance in the development of the model), presence of air bronchogram sign (17.4% importance), smoking status (15.7% importance), types of the lesions (14.8% importance), histology (12.6% importance), presence of pleural indentation sign (7.6% importance), gender (6.9% importance), and presence of lobulation sign (5.6% importance). The ROC analysis achieved an area under the curve (AUC) of 0.854. Multivariate logistic regression analysis demonstrated that this CT-DTA model was an independent predictor of EGFR mutation ($P<0.001$).

Conclusions: CT-DTA model is a simple tool to predict the status of EGFR mutation in SMPLC patients and could be considered for treatment decision-making.

Keywords: Synchronous multiple primary lung cancers (SMPLCs); epidermal growth factor receptor (EGFR); X-ray computed tomography; decision trees

Submitted Sep 21, 2022. Accepted for publication Feb 17, 2023. Published online Mar 09, 2023.
doi: 10.21037/jtd-22-1312
View this article at: https://dx.doi.org/10.21037/jtd-22-1312

Introduction

The widespread screening for lung cancer and recent advances in computed tomography (CT) technology have diagnosed a large number of patients with synchronous multiple primary lung cancers (SMPLCs) in daily practice. Adenocarcinoma accounts for the majority of the reported SMPLC cases (1). Since they are mostly diagnosed at an early stage, the SMPLC patients were divided into three groups based on CT signs as follows: multiple lesions with ground-glass components including both pure ground glass nodules (GGNs) and mixed GGNs (multi-GGN), multiple tumors including one solid nodule plus one or more GGNs (solid-GGN), and multiple solid tumors (multi-solid) (2). Although there is no standard treatment, surgery is the preferred approach for SMPLC patients (3-5). However, considering a patient's performance status, it might be difficult to resect all SMPLC lesions. Especially for medically inoperable patients due to severe comorbidities and limited respiratory function, some non-surgical treatments are available.

Targeted therapy significantly improved the survival rate of lung cancer patients with driver gene mutations (6). But is it feasible for SMPLCs? Typically, driver mutations are highly inconsistent among SMPLCs (7,8). One lesion harboring a driver mutation is not representative of all lesions, making targeted therapy for SMPLCs challenging. With the development of multiple technologies; for example, whole-genome sequencing, a "convergent evolution" scenario was proposed in SMPLCs cases (9). Heterogeneous driver mutations among each lesion from the same patients may converge on the same signaling pathway, such as epidermal growth factor receptor (EGFR) and mitogen-activated protein kinase signaling pathways. Some studies have shown that EGFR mutation occurs frequently in SMPLCs, especially those manifesting as multifocal ground glass opacities (GGOs) (7,8). Thus, a therapeutic option of EGFR-tyrosine kinase inhibitors (TKIs) for EGFR-positive SMPLC patients with inoperable lesions is available. Ye *et al.* proposed a novel strategy involving continual gefitinib treatment for the gefitinib-sensitive lesions and surgical resection for the gefitinib-insensitive lesion (10). Thus, EGFR-TKI therapy presents a potential alternative strategy for SMPLC patients.

Furthermore, biopsy testing has become the gold standard for detecting EGFR mutations. However, limited biopsy specimens, repeated tumor sampling, and a high risk of complications or cancer metastasis can limit the applicability of biopsy. In these situations, a non-invasive and easy-to-use approach is required to predict the EGFR status of SMPLCs.

CT is a routine technique to detect and characterize lung cancer. Previous studies have evaluated the correlation between some CT features and EGFR mutation status in solitary primary lung cancer (11-14). On the other hand, only a few studies have investigated the association of EGFR with CT features for each lesion in SMPLCs (15,16). However, tumor treatment should adhere to the principle of regarding the individual as the dominant factor. Therefore, the present study aimed to develop a people-oriented and CT-based decision tree algorithm model (CT-DTA) model for predicting EGFR mutation status in SMPLCs. We present the following article in accordance with the TRIPOD reporting checklist (available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-1312/rc).

Highlight box

Key findings

- The novel CT-DTA model performed better than traditional individual clinicopathological features and CT signs in predicting EGFR mutation status in synchronous multiple primary lung cancers.
- Synchronous multiple primary lung cancers with EGFR mutations had similar CT findings as single lung adenocarcinoma.

What is known and what is new?

- EGFR mutation occurs frequently in SMPLCs. SMPLC patients at risk of EGFR mutations can be identified based on CT-based decision tree model.

What is the implication, and what should change now?

- CT-DTA model provided the probability to predict EGFR mutation status in synchronous multiple primary lung cancer patients by non-invasive techniques, and could be considered for treatment decision-making.

Methods

Study participants

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This retrospective analysis was approved by the Institutional Review Board of Sun Yat-sen University Cancer Center (IRB No. B2022-293-01), and individual consent for this retrospective analysis was waived. From December 2011 to March 2020, data on consecutive patients with

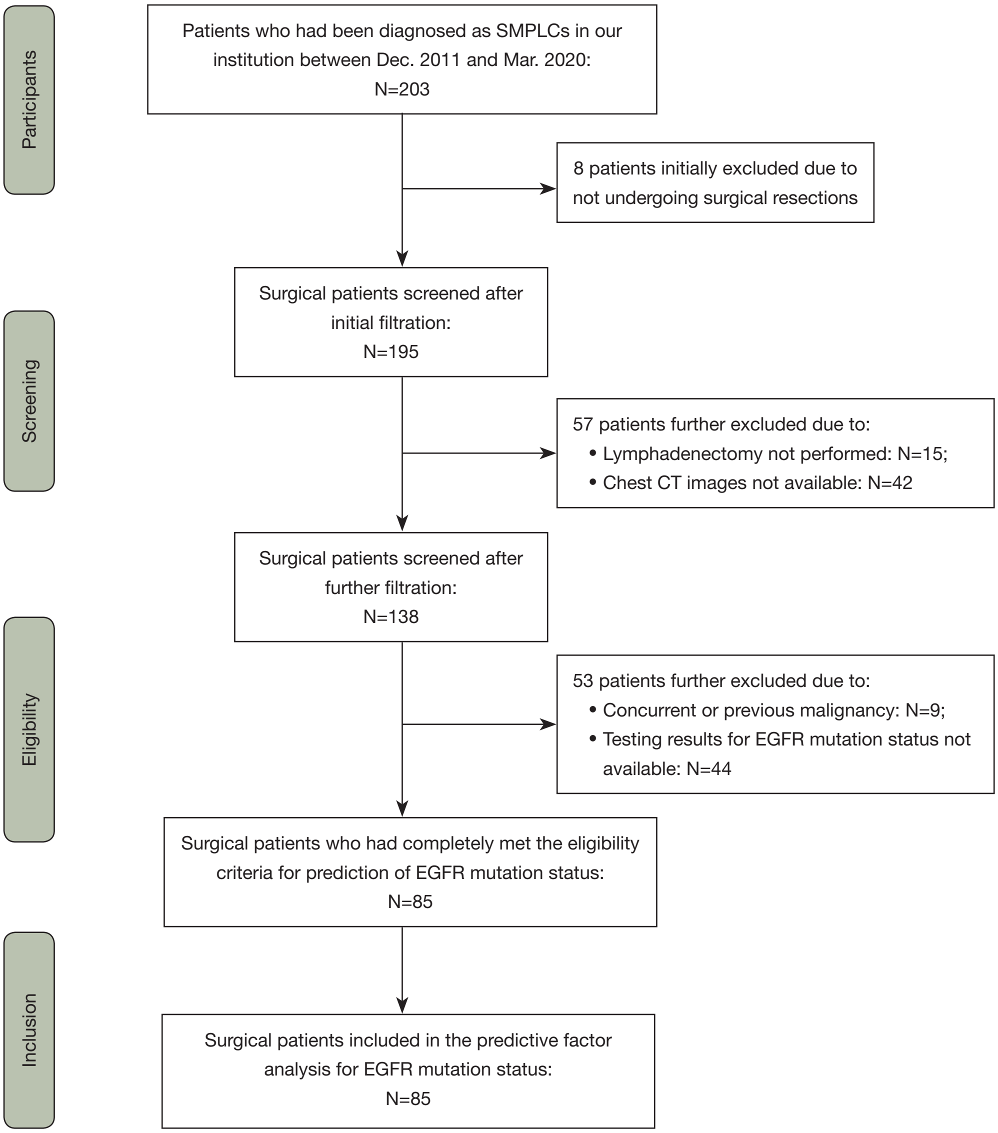
Participants
Patients who had been diagnosed as SMPLCs in our institution between Dec. 2011 and Mar. 2020: N=203
8 patients initially excluded due to not undergoing surgical resections
Screening
Surgical patients screened after initial filtration: N=195
57 patients further excluded due to:
• Lymphadenectomy not performed: N=15;
• Chest CT images not available: N=42
Surgical patients screened after further filtration: N=138
Eligibility
53 patients further excluded due to:
• Concurrent or previous malignancy: N=9;
• Testing results for EGFR mutation status not available: N=44
Surgical patients who had completely met the eligibility criteria for prediction of EGFR mutation status: N=85
Inclusion
Surgical patients included in the predictive factor analysis for EGFR mutation status: N=85

Figure 1 Flowchart of patient selection and exclusion. SMPLC, synchronous multiple primary lung cancer; CT, computed tomography; EGFR, epidermal growth factor receptor.

pulmonary tumors were retrospectively collected in our hospital. According to the 2013 American College of Chest Physicians (ACCP) criteria and the 8th edition lung cancer stage classification (17,18), we identified 203 patients diagnosed with SMPLCs. The inclusion criteria for this study were as follows: synchronous multiple pulmonary tumors resected completely with curative intent, pathologically confirmed by immunohistochemistry (IHC), and results of EGFR mutation status, available clinical data, and thin-section chest CT (slice thickness ≤1 mm) available prior to surgery. The exclusion criteria were as follows: (I) thin-slice chest CT scan was not available; (II) not underwent surgical resections; (III) lymphadenectomy was not performed; (IV) preoperative treatment prior to surgery, such as radiation therapy or chemotherapy; (V) concurrent or previous other malignancies; (VI) testing results for EGFR mutation status not available. Finally, 85 SMPLCs were enrolled in this study (*Figure 1*).

Thin-section CT image analysis

All CT images were reviewed independently by two thoracic radiologists, with both lung window (width, 1,450 HU, level, −500 HU) settings and mediastinal parameters (width, 350 HU; level, 40 HU) in a blinded manner. In case of a disagreement between the two primary radiologists, a third radiologist with 25 years of experience adjudicated a final decision. The following CT characteristics were assessed: (I) number of lesions, lesion types, and locations; (II) lobulation; (III) spiculation; (IV) bubble-like vacuole; (V) air-bronchogram; (VI) pleural indentation; (VII) contour; (VIII) long-axis diameter of the maximal lesion, short-axis diameter of the maximal lesion, and long-axis diameter of the maximal solid portion. The details of the assessment method are described in Table S1.

Clinical data

Patients' clinicopathological characteristics, such as age, gender, smoking status, respiratory comorbidity, cardio-cerebrovascular comorbidity, diabetes mellitus, hepatobiliary comorbidity, family history of malignancy, histology, pleural invasion, lymphovascular invasion, lymph node involvement, and tumor node metastasis (TNM) stage were recorded.

EGFR mutation analysis

EGFR mutations were analyzed through a polymerase chain reaction (PCR)-based amplification refractory mutation system using the Human EGFR Gene Mutations Detection Kit (Roche Diagnostic Products Shanghai Co., Ltd). Based on the outcome of the EGFR test results, the patients were divided into two groups: the EGFR mutation group and the EGFR wild-type group (*Figure 2*).

Statistical analysis

Statistical analyses were performed using IBM SPSS 26.0 software (IBM SPSS Statistics, Version 26.0, Armonk, NY, USA) and R Studio version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria). A statistical significance would be determined by a two-sided test; P value <0.050.

Next, we performed Pearson's chi-square test, Yates's correction test, or Fisher's exact test to assess the association between categorical variables and the EGFR mutation status of SMPLCs. Mann-Whitney U test was conducted to compare the continuous variables between groups of patients. Least absolute shrinkage and selection operator (LASSO) regression was used to select the clinicopathological features and CT signs in two independent regression models related to EGFR mutation to develop a DTA. Using the factors identified in LASSO regression analyses, conditional inference tree analyses were performed to construct a decision tree algorithm (DTA, JMP pro V.14.3, SAS Institute, Cary, NC, USA). The features selected by LASSO regression were input to build a DTA model using the classification and regression trees (CART) algorithm. A putative decision or an action resulted in binary groups. Then, two child nodes were generated from a parent node, and this tree-growing methodology leads to the best split based on the splitting criterion. During this splitting, every child node becomes a parent node. The decision-making process stops when no contribution exists in the further branching. Gini impurity was used for impurity measurement. Receiver operating characteristic (ROC) curve analysis was performed to assess the performance of this DTA model, and the calibration curves were applied to verify the accuracy of the predicted probability for EGFR mutant status by comparison with real-world outcomes (19,20).

The predictive performance of any estimated covariable for EGFR mutation status in SMPLC was elucidated by the multivariate logistic regression analyses, wherein the models were established on the novel CT-based multi-parameter DTA model and other clinicopathological characteristics derived from the LASSO regression analysis. Finally, subgroup analysis on the discriminatory ability of the CT-based multi-parameter DTA model for the EGFR mutant status was conducted using a conditional logistic regression model.

Results

Patient characteristics

The baseline characteristics of the patients are shown in *Table 1*. A total of 85 eligible SMPLC patients (63.6±10.3; range, 57–70 years) were enrolled in this study; among them, 55.3% were women (47/85). The subjects were divided into two groups: the EGFR mutation (n=53) group

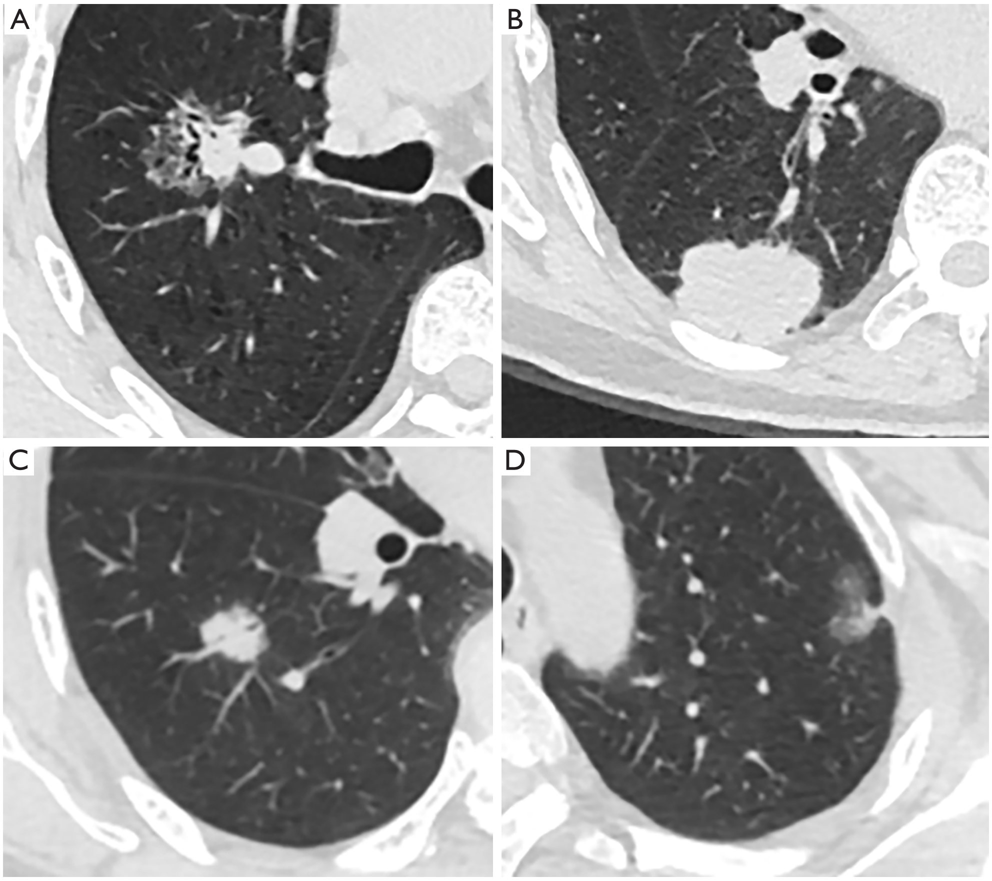

Figure 2 Chest CT images of two SMPLC cases. Patient 1: a 64-year-old male showing a subsolid nodule with EGFR mutation in the right upper lobe (A) and another solid mass with a wild-type mutation in the right lower lobe (B). Patient 2: a 54-year-old female showing a subsolid nodule with 19-Del mutation in the right lower lobe (C) and another subsolid nodule with L858R mutation in the left upper lobe (D). CT, computed tomography; SMPLC, synchronous multiple primary lung cancer; EGFR, epidermal growth factor receptor.

and the EGFR wild-type (n=32) group.

Correlation of EGFR mutation status with clinical and CT features

As shown in *Table 1*, EGFR mutations were detected frequently in females (P=0.003) and those who never smoked (P=0.003). The TNM stage also showed a significant correlation (P=0.034), and 0–I stage was more likely presented as EGFR mutation than other stages. EGFR mutation was frequently seen in adenocarcinoma + adenocarcinoma (AC + AC) group. No significant association was detected between the EGFR mutation and wild-type EGFR groups in terms of respiratory and cardio-cerebrovascular comorbidity, presence of diabetes mellitus, hepatobiliary comorbidity, family history of malignancy, number of lesions, location of the lesions, and pleural and lymphovascular invasion.

Among these patients, 16 cases presented all pure solid nodules (PSNs), 11 cases presented PSN + pure GGO (pGGO), 12 cases presented pGGOs, 21 cases exhibited mixed GGO (mGGO) + PSN, 16 cases as mGGO + pGGO, and 9 cases as all mGGOs. *Table 2* shows the association between CT features and the EGFR mutation status of SMPLCs. The results showed that the types of lesions were associated with EGFR mutation status. No other CT signs were associated with EGFR mutation status.

DTA model

Next, LASSO regression was used to identify some potential predictors related to EGFR mutation among all the

Table 1 Association between clinicopathological characteristics and the EGFR mutation status of SMPLC

Clinicopathological characteristics	Total (N=85)	EGFR mutation status		P value
		Wild (N=32)	Mutant (N=53)	
Age (years)				
Mean ± SD	63.6±10.3	63.3±10.7	63.9±10.2	0.88
Median [IQR]	65 [57–70]	66 [57–72]	65 [57–69]	
Gender, n (%)				
Female	47 (55.3)	11 (34.4)	36 (67.9)	0.003
Male	38 (44.7)	21 (65.6)	17 (32.1)	
Smoking status, n (%)				
Never	59 (69.4)	16 (50.0)	43 (81.1)	0.003
Current/former	26 (30.6)	16 (50.0)	10 (18.9)	
Respiratory comorbidity, n (%)				
Absent	79 (92.9)	29 (90.6)	50 (94.3)	0.67
Present	6 (7.1)	3 (9.4)	3 (5.7)	
Cardio-cerebrovascular comorbidity, n (%)				
Absent	65 (76.5)	26 (81.3)	39 (73.6)	0.42
Present	20 (23.5)	6 (18.8)	14 (26.4)	
Diabetes mellitus, n (%)				
Absent	83 (97.6)	31 (96.9)	52 (98.1)	1.0
Present	2 (2.4)	1 (3.1)	1 (1.9)	
Hepatobiliary comorbidity, n (%)				
Absent	80 (94.1)	31 (96.9)	49 (92.5)	0.65
Present	5 (5.9)	1 (3.1)	4 (7.5)	
Family history of malignancy, n (%)				
Absent	72 (84.7)	26 (81.3)	46 (86.8)	0.71
Present	13 (15.3)	6 (18.8)	7 (13.2)	
Number of the lesions, n (%)				
2	71 (83.5)	26 (81.3)	45 (84.9)	0.89
≥3	14 (16.5)	6 (18.7)	8 (15.1)	
Location of the lesions, n (%)				
Ipsilateral side	53 (62.4)	20 (62.5)	33 (62.3)	0.98
Contralateral side	32 (37.6)	12 (37.5)	20 (37.7)	
Histology, n (%)				
AC + AC	77 (90.6)	26 (81.3)	51 (96.2)	0.048
AC + SCC/SCC + SCC	8 (9.4)	6 (18.8)	2 (3.8)	

Table 1 (*continued*)

Table 1 (*continued*)

Clinicopathological characteristics	Total (N=85)	EGFR mutation status		P value
		Wild (N=32)	Mutant (N=53)	
Pleural invasion, n (%)				
Negative	65 (76.5)	26 (81.3)	39 (73.6)	0.42
Positive	20 (23.5)	6 (18.8)	14 (26.4)	
Lymphovascular invasion, n (%)				
Negative	74 (87.1)	28 (87.5)	46 (86.8)	1.0
Positive	11 (12.9)	4 (12.5)	7 (13.2)	
T stages of the lesions, n (%)				
T_{is-1} + T_{is-1}	47 (55.3)	13 (40.6)	34 (64.2)	0.089
T_{is-1} + T_{2-4}	35 (41.2)	17 (53.1)	18 (34.0)	
T_{2-4} + T_{2-4}	3 (3.5)	2 (6.3)	1 (1.9)	
Lymph node metastasis (N stage), n (%)				
N_0	73 (85.9)	27 (84.4)	46 (86.8)	1.0
N_{1-3}	12 (14.1)	5 (15.6)	7 (13.2)	
TNM stage, n (%)				
0–I	64 (75.3)	20 (62.5)	44 (83.0)	0.034
II	11 (12.9)	8 (25.0%)	3 (5.7)	
III–IV	10 (11.8)	4 (12.5)	6 (11.3)	

EGFR, epidermal growth factor receptor; SMPLC, synchronous multiple primary lung cancer; SD, standard deviation; IQR, interquartile range; AC, adenocarcinoma; SCC, squamous cell carcinoma; TNM, tumor node metastasis.

clinicopathological features and CT signs. It also facilitated variable selection by shrinking down to zero coefficient weights for variables unrelated to EGFR mutation at the optimal log(λ)=−2.5 and −3.5, respectively (*Figure 3*).

Then, a DTA model was developed to predict the EGFR mutant, which had ten binary splits, and eight parameters were used to accurately categorize the lesions as follows: the presence of bubble-like vacuole (19.4% importance in the development of the model), presence of air bronchogram (17.4% importance), smoking status (15.7% importance), types of lesions (14.8% importance), histology (12.6% importance), presence of pleural indentation (7.6% importance), gender (6.9% importance), and presence of lobulation (5.6% importance). The model is illustrated in Table S1 and Figure S1. ROC analysis reached an AUC of 0.854. The calibration curves evaluated the agreement between outcomes predicted by the DTA model and the real-world outcomes (*Figure 4*).

Further evaluation of the CT-DTA model

To elucidate the performance of the CT-DTA compared to other traditional clinicopathological features and CT signs, we constructed two multivariate binary logistic regression models. One was established on the original parameters estimated on chest CT images and other clinicopathological characteristics derived from the LASSO regression analysis (Hosmer-Lemeshow test P=0.46); the other was established on the novel CT-based multi-parameter DTA model and other clinicopathological characteristics derived from the LASSO regression analysis (Hosmer-Lemeshow test P=0.76). As shown in *Table 3*, we found that the CT-DTA was an independent predictor of EGFR mutation (P<0.001).

Subgroup analysis

The results of the subgroup analysis on the discriminatory

Table 2 Association between CT-detected parameters and the EGFR mutation status of SMPLC

CT-detected parameters	Total (N=85)	EGFR mutation status		P value
		Wild (N=32)	Mutant (N=53)	
Types of the lesions, n (%)				
All PSNs	16 (18.8)	10 (31.3)	6 (11.3)	0.037
PSN + pGGO	11 (13.0)	5 (15.7)	6 (11.3)	
All pGGOs	12 (14.2)	6 (18.8)	6 (11.3)	
mGGO + PSN	21 (24.7)	6 (18.8)	15 (28.3)	
mGGO + pGGO	16 (18.8)	4 (12.5)	12 (22.6)	
All mGGOs	9 (10.6)	1 (3.1)	8 (15.1)	
Spiculation, n (%)				
Absent	41 (48.2)	16 (50.0)	25 (47.2)	0.88
1 lesion present	32 (37.6)	11 (34.4)	21 (39.6)	
≥2 lesions present	12 (14.1)	5 (15.6)	7 (13.2)	
Lobulation, n (%)				
Absent	5 (5.9)	4 (12.5)	1 (1.9)	0.13
1 lesion present	21 (24.7)	7 (21.9)	14 (26.4)	
≥2 lesions present	59 (69.4)	21 (65.6)	38 (71.7)	
Bubble-like vacuole, n (%)				
Absent	55 (64.7)	24 (75.0)	31 (58.5)	0.11
1 lesion present	27 (31.8)	8 (25.0)	19 (35.8)	
≥2 lesions present	3 (3.5)	0 (0%)	3 (5.7)	
Air Bronchogram, n (%)				
Absent	40 (47.1)	20 (62.5)	20 (37.7)	0.076
Normally present	27 (31.8)	8 (25.0)	19 (35.8)	
≥1 lesion pathologically present	18 (21.2)	4 (12.5)	14 (26.4)	
Pleural indentation, n (%)				
Absent	21 (24.7)	12 (37.5)	9 (17.0)	0.096
1 lesion present	49 (57.6)	16 (50.0)	33 (62.3)	
≥2 lesions present	15 (17.6)	4 (12.5)	11 (20.8)	
Contour, n (%)				
Round/oval + round/oval	78 (91.8)	28 (87.5)	50 (94.3)	0.21
Round/oval + irregular	6 (7.1)	4 (12.5)	2 (3.8)	
Irregular + irregular	1 (1.2)	0 (0.0)	1 (1.9)	
Long-axis diameter of the maximal lesion (mm)				
Mean ± SD	27.3±14.0	28.2±15.7	26.8±12.9	0.79
Median [IQR]	27 [18–36]	27 [18–40]	27 [19–31]	
Short-axis diameter of the maximal lesion (mm)				
Mean ± SD	21.6±11.1	22.9±12.5	20.8±10.3	0.49
Median [IQR]	20 [13–28]	22 [13–34]	19 [14–25]	
Long-axis diameter of the maximal solid portion (mm)				
Mean ± SD	22.9±16.7	24.3±18.3	22.0±15.7	0.54
Median [IQR]	22 [13–32]	22 [12–38]	21 [13–30]	

CT, computed tomography; EGFR, epidermal growth factor receptor; SMPLC, synchronous multiple primary lung cancer; PSN, pure solid nodule; pGGO, pure ground glass opacity; mGGO, mixed ground glass opacity; SD, standard deviation; IQR, interquartile range.

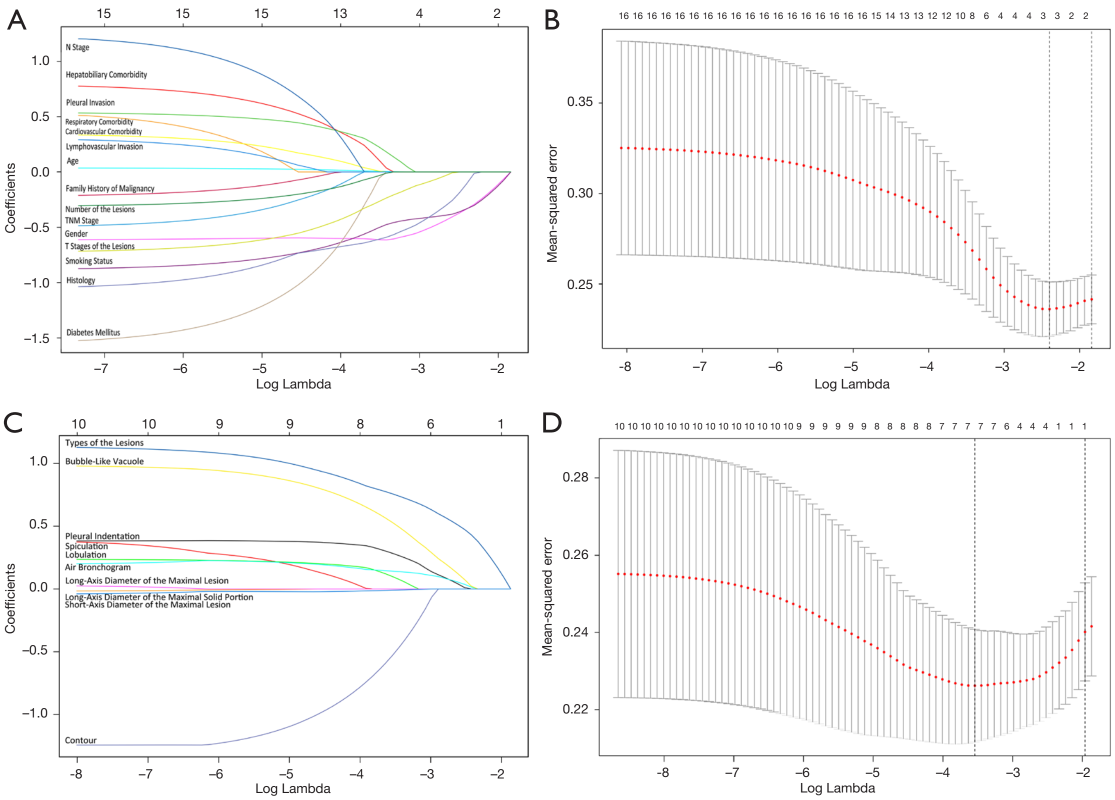

Figure 3 LASSO coefficient profiles of the clinicopathological features (A) and CT signs (C), respectively. Dotted vertical lines are drawn at the optimal values using the minimum criteria and the 1 standard error of the minimum criteria (the 1-SE criteria) using the optimal log(λ) =−2.5 and −3.5, respectively (B,D). TNM, tumor node metastasis; LASSO, least absolute shrinkage and selection operator; CT, computed tomography; SE, standard error.

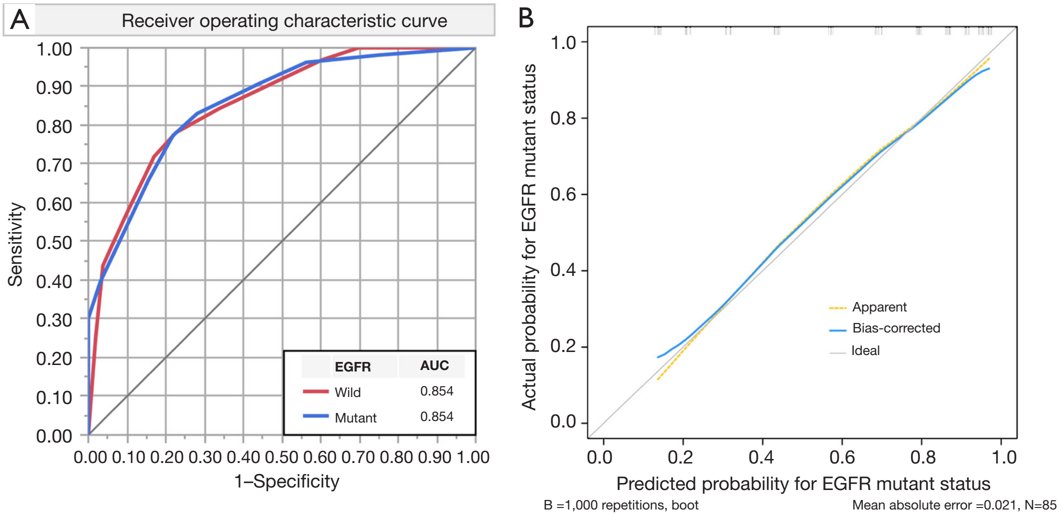

Figure 4 The ROC curves and calibration plot of CT-based DTA model. (A) ROC analysis reached an AUC of 0.854. (B) The 45° dashed line illustrates the ideal prediction that the actual probability is equal to the predicted probability. The plot represents the accuracy of the best-fit model ("Apparent") and the bootstrap model ("Bias-corrected") for predicting the EGFR mutations. EGFR, epidermal growth factor receptor; AUC, area under the curve; ROC, receiver operating characteristic; CT, computed tomography; DTA, decision tree algorithm.

Table 3 Multivariable logistic regression analyses of predictive factors for EGFR mutation status in SMPLC

Characteristics	Multivariable analysis^A		Multivariable analysis^B	
	OR (95% CI)	P value	OR (95% CI)	P value
Gender (male *vs.* female)	0.63 (0.12–3.19)	0.57	–	–
Smoking status (current/former *vs.* never)	0.20 (0.027–1.42)	0.11	–	–
Histology (AC + SCC/SCC + SCC *vs.* AC + AC)	0.17 (0.009–2.94)	0.22	–	–
Types of the lesions (≥1 mGGOs *vs.* all pGGOs/ pGGO + PSN *vs.* all PSNs)	1.20 (0.70–2.08)	0.51	–	–
Presence of lobulation (≥2 lesions *vs.* 1 lesion *vs.* absent)	1.74 (0.61–4.94)	0.30	–	–
Presence of bubble-like vacuole (≥2 lesions *vs.* 1 lesion *vs.* absent)	3.32 (0.94–11.76)	0.063	–	–
Presence of air bronchogram (pathologically *vs.* normally *vs.* absent)	1.47 (0.66–3.26)	0.35	–	–
Presence of pleural indentation (≥2 lesions *vs.* 1 lesion *vs.* absent)	1.88 (0.72–4.92)	0.20	–	–
Contour (round/oval *vs.* irregular)	0.22 (0.032–1.46)	0.12	0.21 (0.038–1.12)	0.067
Short-axis diameter of the maximal lesion (mm) (per 1 mm increased)	1.030 (0.958–1.106)	0.43	1.010 (0.960–1.062)	0.71
CT-based multi-parameter decision tree algorithm model (per split proceed)	–	–	1.80 (1.40–2.32)	<0.001

^{A}, the multivariable binary logistic regression model was established on the original parameters estimated on chest CT images and other clinicopathological characteristics derived from the LASSO regression analysis (Hosmer-Lemeshow test P=0.46); ^{B}, the multivariable binary Logistic regression model was established on the novel CT-based multi-parameter decision tree algorithm model and other clinicopathological characteristics derived from the LASSO regression analysis (Hosmer-Lemeshow test P=0.76). EGFR, epidermal growth factor receptor; SMPLC, synchronous multiple primary lung cancer; AC, adenocarcinoma; SCC, squamous cell carcinoma; mGGO, mixed ground glass opacity; pGGO, pure ground glass opacity; PSN, pure solid nodule; CT, computed tomography; OR, odds ratio; CI, confidence interval; LASSO, least absolute shrinkage and selection operator.

ability of the CT-based multi-parameter DTA model for the EGFR mutant status are shown in *Table 4*. Forest plot depicting the hazard ratios is shown in *Figure 5*. The CT-based multi-parameter DTA model showed a good performance in subgroups categorized according to the patients' age (<65/≥65 years), gender (female/male), smoking status (never/current or former), location of the lesions (ipsilateral/contralateral), histology (AC + AC), T stage of the lesions (Tis-1 + Tis-1/T2-4 + Tis-1 and T2-4 + T2-4), lymph node metastasis (N0 stage), and TNM stage (0–I/II–IV).

Discussion

In patients with multiple pulmonary sites of involvement, distinguishing between multiple primary lung cancers (MPLCs) and intrapulmonary metastasis (IPM) is critical for developing a therapeutic strategy. Suh *et al.* applied one algorithm based on comprehensive information on clinical and imaging variables that allows differentiation between MPLCs and IPMs. Furthermore, predicting the status of EGFR mutation in MPLCs might facilitate personalized precision treatment of these patients (21). In the current study, the novel CT-based multi-parameter DTA model performed better than traditional individual clinicopathological features and CT signs in predicting the EGFR mutation status. This DTA model to predict EGFR mutant had ten binary splits; we used three clinicopathological parameters and five CT features to accurately categorize the lesions and also found that the DTA model was an independent predictor of EGFR mutation. To the best of our knowledge, this is the first description of a multiparametric prediction model for predicting the EGFR mutation status in SMPLC, which

Table 4 Subgroup analyses on the discriminatory ability of the CT-based multi-parameter decision tree algorithm model for the EGFR mutant status

Subgroups stratified	AUC	95% CI	P value
Age (years)			
<65	0.808	0.671–0.946	0.002
≥65	0.880	0.782–0.979	<0.001
Gender			
Female	0.813	0.686–0.940	0.002
Male	0.836	0.702–0.970	<0.001
Smoking status			
Never	0.851	0.746–0.956	<0.001
Current & former	0.756	0.558–0.955	0.031
Location of lesions			
Ipsilateral	0.867	0.770–0.963	<0.001
Contralateral	0.854	0.717–0.992	0.001
Histology			
AC + AC	0.834	0.741–0.926	<0.001
AC + SCC/SCC + SCC	0.750	0.267–1.000	0.32
T stages of the lesions			
T_{is-1} + T_{is-1}	0.760	0.613–0.907	0.006
T_{2-4} + $T_{is-1/}T_{2-4}$ + T_{2-4}	0.910	0.820–1.000	<0.001
Lymph node metastasis (N stage)			
N_0	0.840	0.749–0.931	<0.001
N_{1-3}	1.000	1.000–1.000	0.004
TNM stage			
0–I	0.824	0.719–0.928	<0.001
II–IV	0.921	0.810–1.000	0.001

CT, computed tomography; EGFR, epidermal growth factor receptor; AUC, area under curve; AC, adenocarcinoma; SCC, squamous cell carcinoma; TNM, tumor node metastasis; CI, confidence interval.

can be easily applied in clinical practice.

Next, we demonstrated that the first step is to identify the genders. In this study, EGFR mutations were frequently detected in women and in those who never smoked; these results were consistent with those reported previously (15,22). However, tumor EGFR mutations were detected in patients with clinical characteristics other than female sex, adenocarcinoma histology, or never-smoking status (23). In a retrospective study of 2,142 lung adenocarcinomas, EGFR mutations from men represented 31% of the occurrences (24). If the gender was male and the histology of SMPLC indicated adenocarcinoma, the presence of an air bronchogram was a risk factor for EGFR mutations. Previous studies have reported that air bronchogram was associated with pathological invasiveness and activated EGFR mutations (25-27). The classification of the air bronchogram is diverse; however, abnormalities (dilated or tortuous bronchi and abruptly obstructed bronchi) were observed frequently with increasing invasiveness. In the current study, men with multiple primary lung adenocarcinomas are likely to present EGFR mutations if they have abnormal air bronchogram signs on CT. When

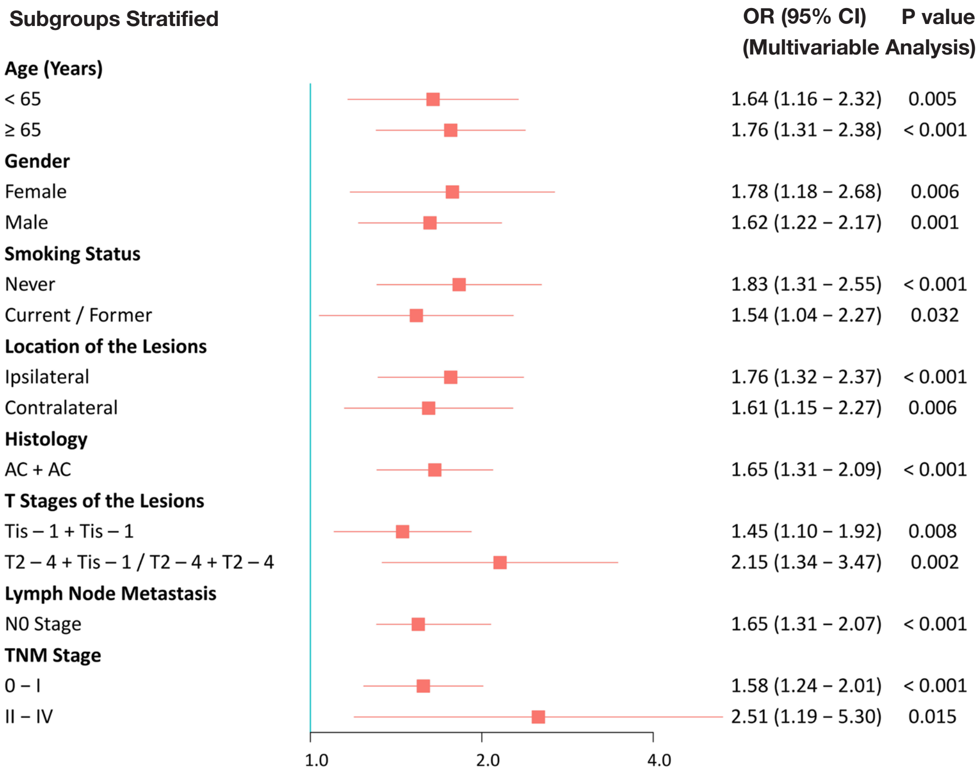

Figure 5 Forest plot of the subgroup analysis on the discriminatory ability of the CT-based multi-parameter DTA model for the EGFR mutant status. AC, adenocarcinoma; TNM, tumor node metastasis; OR, odds ratio; CI, confidence interval; CT, computed tomography; DTA, decision tree algorithm; EGFR, epidermal growth factor receptor.

the CT finding of the air bronchogram was absent or normally present, the EGFR mutation could be related to the next branch of the decision tree concerning the presence of pleural indentation and smoking status.

Moreover, if the gender was female, the combination of pleural indentation (≥1 lesion), presence of lobulation (≥2 lesions), presence of bubble-like vacuole sign, and presence of GGO on CT suggested that tumors in SMPLC were susceptible to EGFR mutations. The importance of each feature in the creation of the DTA model for predicting the EGFR mutation status of SMPLC patients is graphically demonstrated in Table S1 and Figure S1. The CT sign of bubble-like vacuole, air bronchogram sign, types of lesions, smoking status, and histology accounted for >10% importance in the tree creation. These findings differ slightly from those reported previously (15,16). Some investigators also identified the correlation between EGFR mutations and CT characteristics in multiple primary lung adenocarcinoma (MPLA) patients. However, their results showed that only GGO was correlated with EGFR mutation, and there was no correlation between CT features and EGFR mutations (15). Although many studies analyzed the correlation between CT characteristics and EGFR mutation in single non-small cell lung cancer (NSCLC), the results are yet controversial. A recent meta-analysis of 17 original studies provided evidence of an association between CT features and EGFR mutation in a single NSCLC (13). These differences could be attributed to the different study designs and research variables. Based on these results, it could be deduced that SMPLC with EGFR mutations had CT findings similar to those of single NSCLC, which confirms that SMPLC patients at risk of EGFR mutations can be identified based on CT features.

To the best of our knowledge, this is the first study to build a CT-DTA model for predicting EGFR mutation status in SMPLC patients, and ROC analysis reached an AUC of 0.854. The results of our study are significantly superior to the previous study by Han *et al.* (15). Moreover, the DTA tree model has strong explanatory power and provides a visual representation of the decision-making process for easy guidance and application.

Nevertheless, the present study has several limitations. First, this was a retrospective study limited to SMPLC patients who underwent surgery, with the inherent possibility of case selection bias. Second, we used a consensus reading for CT interpretation; hence, inter-reader correlation could not be assessed. Third, this study lacked external validation that should be carried out in a large independent population. Nonetheless, according to our subgroup analysis, this simple CT-DTA model had good discrimination for outcome prediction and could be easily applied in clinics.

Conclusions

In conclusion, this study presented a CT-based decision tree model to predict the status of EGFR mutation in multiple synchronous lung cancers, which might facilitate personalized precision treatment of patients.

Acknowledgments

We would like to thank Sunil Saroj for his help in polishing our paper.
Funding: This work was supported by Shanghai Municipal Population and Family Planning Commission (Grant No. 20184Y0037).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-1312/rc

Data Sharing Statement: Available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-1312/dss

Peer Review File: Available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-1312/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-1312/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This retrospective analysis was approved by the Institutional Review Board of Sun Yat-sen University Cancer Center (IRB No. B2022-293-01), and individual consent for this retrospective analysis was waived.

References

1. Chang YL, Wu CT, Lee YC. Surgical treatment of synchronous multiple primary lung cancers: experience of 92 patients. J Thorac Cardiovasc Surg 2007;134:630-7.
2. Zhang Y, Li G, Li Y, et al. Imaging Features Suggestive of Multiple Primary Lung Adenocarcinomas. Ann Surg Oncol 2020;27:2061-70.
3. Chen TF, Xie CY, Rao BY, et al. Surgical treatment to multiple primary lung cancer patients: a systematic review and meta-analysis. BMC Surg 2019;19:185.
4. Chen C, Huang X, Peng M, et al. Multiple primary lung cancer: a rising challenge. J Thorac Dis 2019;11:S523-36.
5. Zhao L, Liu C, Xie G, et al. Multiple Primary Lung Cancers: A New Challenge in the Era of Precision Medicine. Cancer Manag Res 2020;12:10361-74.
6. Arbour KC, Riely GJ. Systemic therapy for locally advanced and metastatic non-small cell lung cancer: a review. JAMA 2019;322:764-74.
7. Wu C, Zhao C, Yang Y, et al. High Discrepancy of Driver Mutations in Patients with NSCLC and Synchronous Multiple Lung Ground-Glass Nodules. J Thorac Oncol 2015;10:778-83.
8. Liu M, He WX, Song N, et al. Discrepancy of epidermal growth factor receptor mutation in lung adenocarcinoma presenting as multiple ground-glass opacities. Eur J Cardiothorac Surg 2016;50:909-13.
9. Ma P, Fu Y, Cai MC, et al. Simultaneous evolutionary expansion and constraint of genomic heterogeneity in multifocal lung cancer. Nat Commun 2017;8:823.
10. Ye C, Wang J, Li W, et al. Novel Strategy for Synchronous

Multiple Primary Lung Cancer Displaying Unique Molecular Profiles. Ann Thorac Surg 2016;101:e45-7.
11. Liu Y, Kim J, Qu F, et al. CT Features Associated with Epidermal Growth Factor Receptor Mutation Status in Patients with Lung Adenocarcinoma. Radiology 2016;280:271-80.
12. Yano M, Sasaki H, Kobayashi Y, et al. Epidermal growth factor receptor gene mutation and computed tomographic findings in peripheral pulmonary adenocarcinoma. J Thorac Oncol 2006;1:413-6.
13. Zhang H, Cai W, Wang Y, et al. CT and clinical characteristics that predict risk of EGFR mutation in non-small cell lung cancer: a systematic review and meta-analysis. Int J Clin Oncol 2019;24:649-59.
14. Cheng Z, Shan F, Yang Y, et al. CT characteristics of non-small cell lung cancer with epidermal growth factor receptor mutation: a systematic review and meta-analysis. BMC Med Imaging 2017;17:5.
15. Han X, Fan J, Gu J, et al. CT features associated with EGFR mutations and ALK positivity in patients with multiple primary lung adenocarcinomas. Cancer Imaging 2020;20:51.
16. Huo JW, Luo TY, He XQ, et al. Radiological classification, gene-mutation status, and surgical prognosis of synchronous multiple primary lung cancer. Eur Radiol 2022;32:4264-74.
17. Kozower BD, Larner JM, Detterbeck FC, et al. Special treatment issues in non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013;143:e369S-99S.
18. Detterbeck FC, Nicholson AG, Franklin WA, et al. The IASLC Lung Cancer Staging Project: Summary of Proposals for Revisions of the Classification of Lung Cancers with Multiple Pulmonary Sites of Involvement in the Forthcoming Eighth Edition of the TNM Classification. J Thorac Oncol 2016;11:639-50.
19. Kha QH, Ho QT, Le NQK. Identifying SNARE Proteins Using an Alignment-Free Method Based on Multiscan Convolutional Neural Network and PSSM Profiles. J Chem Inf Model 2022;62:4820-6.
20. Lam LHT, Do DT, Diep DTN, et al. Molecular subtype classification of low-grade gliomas using magnetic resonance imaging-based radiomics and machine learning. NMR Biomed 2022;35:e4792.
21. Suh YJ, Lee HJ, Sung P, et al. A Novel Algorithm to Differentiate Between Multiple Primary Lung Cancers and Intrapulmonary Metastasis in Multiple Lung Cancers With Multiple Pulmonary Sites of Involvement. J Thorac Oncol 2020;15:203-15.
22. Izumi M, Oyanagi J, Sawa K, et al. Mutational landscape of multiple primary lung cancers and its correlation with non-intrinsic risk factors. Sci Rep 2021;11:5680.
23. Shi Y, Au JS, Thongprasert S, et al. 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 2014;9:154-62.
24. D'Angelo SP, Pietanza MC, Johnson ML, et al. Incidence of EGFR exon 19 deletions and L858R in tumor specimens from men and cigarette smokers with lung adenocarcinomas. J Clin Oncol 2011;29:2066-70.
25. Dai J, Shi J, Soodeen-Lalloo AK, et al. Air bronchogram: A potential indicator of epidermal growth factor receptor mutation in pulmonary subsolid nodules. Lung Cancer 2016;98:22-8.
26. Han X, Fan J, Li Y, et al. Value of CT features for predicting EGFR mutations and ALK positivity in patients with lung adenocarcinoma. Sci Rep 2021;11:5679.
27. Zhang Y, Qiang JW, Shen Y, et al. Using air bronchograms on multi-detector CT to predict the invasiveness of small lung adenocarcinoma. Eur J Radiol 2016;85:571-7.

Cite this article as: Luo Y, Li S, Ma H, Zhang W, Liu B, Xie C, Li Q. CT-based decision tree model for predicting EGFR mutation status in synchronous multiple primary lung cancers. J Thorac Dis 2023;15(3):1196-1209. doi: 10.21037/jtd-22-1312

Supplementary

Table S1 CT characteristics of synchronous multiple primary lung cancers

Characteristics	Definition
Type of lesions	(I) Pure ground glass opacity: tumor without a solid component that obscures the underlying lung parenchyma other than blood vessels on thin-slice CT viewed on CT lung window settings (width, 1,450 HU; level,−500HU) (II) Mixed ground glass opacity: tumor with a solid component obscuring the underlying lung parenchyma other than blood vessels on thin-slice CT scan viewed on CT lung window settings (III) Pure solid nodule
Lobulation	Tumor's surface showed a wavy or scalloped configuration
Spiculation	Lines radiating from the margins of the tumor
Bubble-like vacuole	The presence of air in the tumor
Air-bronchogram	Tube-like or branched air structure within the tumor
Pleural indentation	Tumor adheres to the pleura or fissure, and the pleura indentation with one or more stripes
Contour	The overall shape of lesion
Long-axis diameter of the maximal lesion	Longest diameter of the larger tumor on lung window setting
Short-axis diameter of the maximal lesion	Longest perpendicular diameter in the same section of the larger tumor on lung window setting
Long-axis diameter of the maximal solid portion	Longest diameter of the largest solid component measured on lung window setting

https://dx.doi.org/10.21037/jtd-22-1312

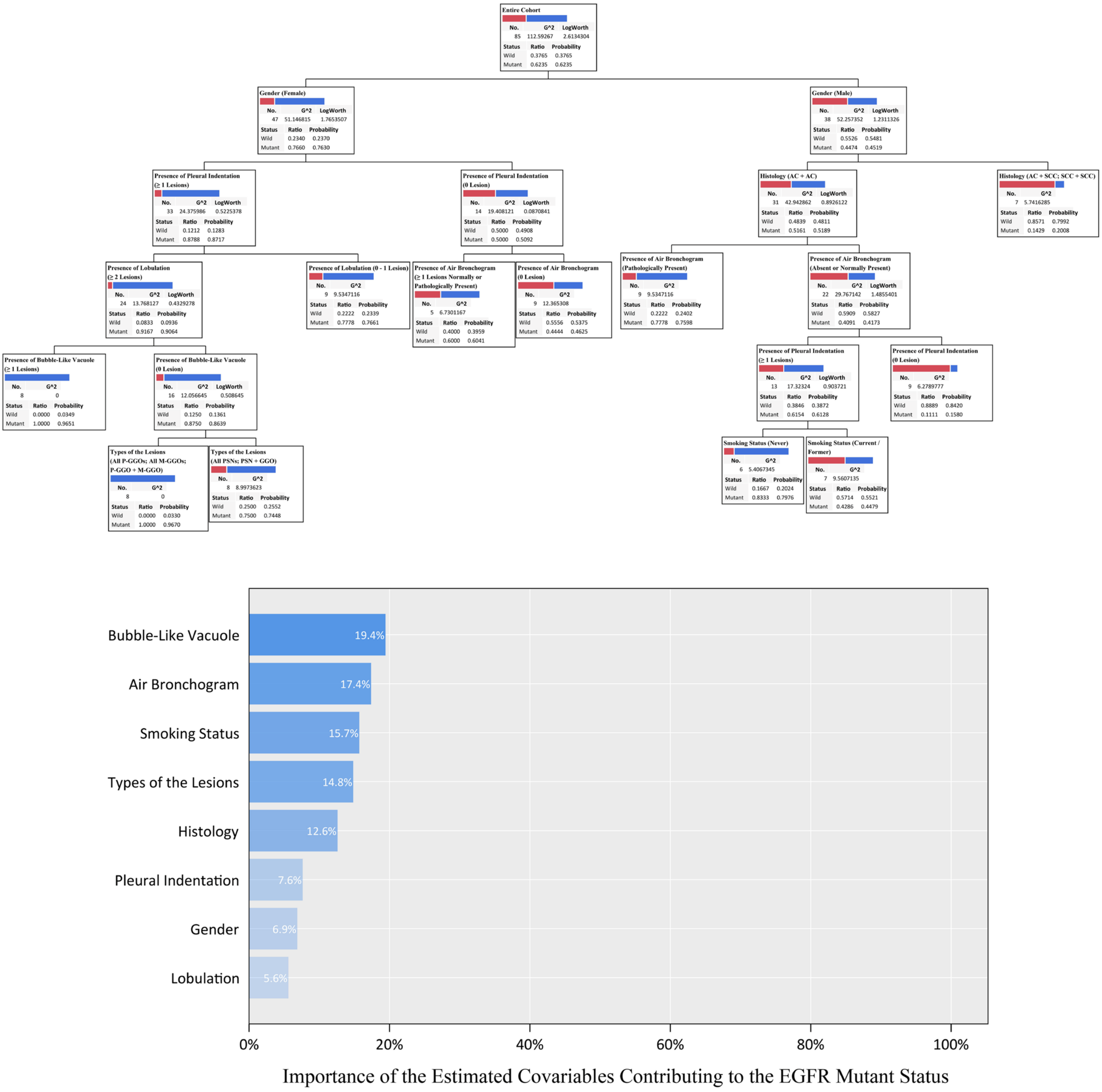

Figure S1 This decision tree model had ten binary splits and used eight parameters to accurately categorize lesions, the importance of each feature in the development of the DTA model is graphically demonstrated.

https://dx.doi.org/10.21037/jtd-22-1312