CT and CEA-based machine learning model for predicting malignant pulmonary nodules

Man Liu1,2 | Zhigang Zhou3 | Fenghui Liu1 | Meng Wang3 | Yulin Wang2 | Mengyu Gao3 | Huifang Sun3 | Xue Zhang2 | Ting Yang2,4 | Longtao Ji2,4 | Jiaqi Li2 | Qiufang Si2,4 | Liping Dai2,4 | Songyun Ouyang1

1Department of Respiratory and Sleep Medicine, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
2Department of Respiratory and Sleep Medicine, Henan Institute of Medical and Pharmaceutical Sciences, Zhengzhou, China
3Department of Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
4BGI College, Zhengzhou University, Zhengzhou, China

Abstract
Computed tomography (CT), an efficient radiological technology, is used to detect lung cancer in the clinic. Carcinoembryonic antigen (CEA), a common tumor biomarker, is applied in the detection of various tumors. To highlight the advantages of two-dimensional techniques and assist clinicians in optimizing lung cancer diagnostic schemes, we established a favorable model combining CT and CEA. In the study, univariate analysis was performed to screen independent predictors in a training cohort of 271 patients with malignant pulmonary nodules (MPNs) and 92 with benign pulmonary nodules (BPNs). Six machine learning–based models involving five CT predictors (mediastinal lymph node enlargement, lobulation, vascular notch sign, spiculation, and nodule number) and lnCEA were constructed and validated in an independent cohort of 129 participants (92 MPNs and 37 BPNs) by SPSS Modeler. A nomogram and the Delong test were generated by R software. Finally, the model established by logistic regression had highest diagnostic efficiency (area under the curve [AUC] = 0.912). Moreover, the diagnostic ability of the logistic model in the validation cohort (AUC = 0.882, 80.4% sensitivity, 75.7% specificity) was higher than that of the Peking University model (AUC = 0.712, 68.5% sensitivity, 70.3% specificity) and the Mayo model (AUC = 0.745, 62.0% sensitivity, 75.7% specificity). Interestingly, for the participants with intermediate (10-30 mm) and CEA-negative nodule, the model reached an AUC of 0.835 (72.3% sensitivity, 83.3% specificity). The AUC for the early...
lungs were as high as 0.822 with 67.3% sensitivity and 78.9% specificity. As a conclusion, this promising model presents a new diagnostic strategy for the clinic to distinguish MPNs from BPNs.

**KEYWORDS**

BPNs, CEA, CT, logistic model, MPNs

## 1 | INTRODUCTION

Lung cancer is the leading cause of cancer-related death in men and women. As early symptoms are atypical, patients tend to be diagnosed in advanced stage. The 5-year survival of distant-stage lung cancer is only 5%, whereas it could be up to 57% for localized stage. Therefore, an effective and reliable means for the diagnosis of lung cancer, especially early lung cancer, could contribute to the improvement of the prognosis and survival of patients.

Due to its accuracy in lung cancer diagnosis, computed tomography (CT) is widely used in the clinic. However, CT has a certain false positive rate (FPR) in screening lung cancer, which increases financial costs and psychological pressure for the patients. Moreover, repeated CT examination increases the risk of lung cancer.

Therefore, reducing the false positive value (FPV) of CT and developing a method to distinguish malignant pulmonary nodules (MPNs) from benign pulmonary nodules (BPNs) could promote the application of CT in the clinic and help exclude BPNs from subsequent unnecessary treatment.

Carcinoembryonic antigen (CEA), a 180-kD glycoprotein, was first described in 1965 by Gold and Freedman. CEA is regarded as a human tumor biomarker, especially for gastrointestinal tumors. At the same time, serum CEA is widely used for progression monitoring of lung cancer. Moreover, several research studies have reported that serum CEA had a high specificity in detecting lung cancer. While looking for new preferable diagnostic biomarkers, traditional biomarkers in the clinic should be reasonably applied to support the diagnosis of lung cancer.

Nowadays, multiple models are used for the differential diagnosis of pulmonary nodules (PNs), such as the Mayo model and Peking University model. However, various potential issues exist; for example, instability, single index, and small sample size. In the present study, we aimed to establish a model combining CT and CEA to make use of the mutual advantage. We collected the CEA concentration and 14 CT characteristics of the 363 participants from the training data set. Univariate and multivariate analyses were used to identify the variables which could distinguish MPNs from BPNs. Based on the candidate factors, we developed six different algorithmic models combining CEA and CT via SPSS Modeler in the training data set and validated them in an independent data set to differentiate MPNs and BPNs. Subsequently, we compared the optimal model with the Peking University model and Mayo model in the validation data set. Finally, we evaluated the differential diagnostic ability of the selected model for MPNs and BPNs in stratified analyses.

## 2 | MATERIALS AND METHODS

### 2.1 | Study population

A total of 492 participants who met the inclusion criteria were enrolled from the First Affiliated Hospital of Zhengzhou University. The inclusion criteria were as follows: (1) CT and serum CEA detection of the participants were performed in the first Affiliated Hospital of Zhengzhou University, and the clinical data were available; (2) patients with PNs were newly confirmed by histopathologic diagnosis, and history of malignancy was not found in patients with BPNs; (3) participants were all older than 18 years. The exclusion criteria contained the following terms: (1) participants without complete clinical data and clear pathological results; (2) not newly diagnosed PNs and patients with BPNs having a history of malignancy; (3) participants younger than 18 years of age. The workflow of study design and the clinical information of participants are displayed in Figure 1 and Table 1, respectively. The study was approved by the Medical Ethics Committee of the First Affiliated Hospital of Zhengzhou University (2021-KY-1057-002).

1. Training data set

A total of 363 participants in the training data set were recruited from September 2016 to March 2019, whose information constituted the sample library of PNs. The training data set was used to construct the diagnostic model based on the logistic regression for the differentiation between the MPNs and BPNs.

2. Validation data set

The diagnostic performance of the logistic model was validated in an independent data set of 129 participants who underwent lung puncture biopsy in January 2018.

### 2.2 | Electrochemiluminescence and CT detection

Serum CEA concentrations of all participants were measured by electrochemiluminescence on an automatic analyzer operated by professional technicians of the First Affiliated Hospital of Zhengzhou University. The CEA values were distributed widely (0.12-5311 ng/ml), and in order to narrow the data range, eliminate heteroscedasticity, and make the data stable, we converted CEA to lnCEA via the transformation algorithm. The transformation could not affect the nature and correlation of the data.
Spiral CT images were acquired using a 64-detector CT row scanner. The CT data were stored in the imaging archive and communication system. The 14 nodule characteristics data of CT (diameter, number, type, morphology, margin, cavity, spiculation, vascular notch sign, lobulation, spines, pleural indentation, calcification, emphysema, and mediastinal lymph node enlargement) were rated by a senior radiologists using a 2-megapixel gray-scale monitor at lung windows and mediastinal windows.

### 2.3 Statistical analysis

Data analysis was performed by SPSS (Version 21.0) and GraphPad Prism software (version 6.0). The chi-squared test and the Mann-Whitney U test were applied for comparing the variable values between MPNs and BPNs. Six machine learning–based models (logistic, multilayer perceptron [MLP], radial basis function network [RBF], C5.0, chi-squared automatic interaction detection [CHAID], and support vector machines [SVM]) were employed to develop the predictive models for MPNs in the training data set and validated in an independent data set via SPSS Modeler 18.0 software. Nomogram production and DeLong test were generated by R4.0.3. Receiver-operating characteristic (ROC) curve analysis and area under the curve (AUC) were applied to evaluate the diagnostic value of each variable or model. The sensitivity and specificity were decided by the maximum Youden index. In all tests, $p < 0.05$ (two-sided) was considered significant.

## 3 RESULTS

### 3.1 Patient characteristics

A total of 492 participants with definitive diagnosis were included in the training ($n = 363$) and validation ($n = 129$) data sets. The clinical information of the participants is summarized in Table 1.

The distribution of benign and malignant nodules in the training and validation data sets according to lesion size (small: <10 mm, intermediate: 10-30 mm, and large: ≥30 mm) are displayed in Figure S1. The small nodules were mostly BPNs, while the large ones tended to be MPNs. The proportion of BPNs and MPNs in the intermediate (10-30 mm) nodules was similar (Figure S1).

### 3.2 Characteristic selection

According to the univariate and ROC analyses, the potential indicators for MPNs included lnCEA, diameter, number, type, morphology, margi...
margin, cavity, spiculation, vascular notch sign, lobulation, spines, pleural indentation, calcification, emphysema, and mediastinal lymph node enlargement (Table S1). Six indicators (lnCEA, number, spiculation, vascular notch sign, lobulation, and mediastinal lymph node enlargement) were identified as independent diagnostic biomarkers for MPNs by multivariate logistic regression analysis (Table S1 and Figure 2A–F). The AUC range of six predictors for distinguishing MPNs from BPNs ranged from 0.625 to 0.783. Among them, lobulation had the highest diagnostic value with an AUC of 0.783.

### 3.3 Model construction and validation

Four machine learning algorithms (logistic regression, artificial neural networks [ANN], decision trees [DTs], and SVM) were used to construct six models (logistic, MLP, RBF, C5.0, CHAID, and SVM) via SPSS modeler 18.0. Among them, the MLP and RBF models were established based on ANN, while the CHAID and C5.0 models were DTs. Models (logistic, MLP, RBF, CHAID and SVM) included six indicators (lnCEA, number, spiculation, vascular notch sign, lobulation, and mediastinal lymph node enlargement), while model C5.0 included five indicators except mediastinal lymph node enlargement. After comparison, the logistic model had the highest diagnostic efficiency (Table 2 and Figure 3A,B). The DeLong test showed that in the AUCs of the logistic model there was no significant statistical difference between the training and validation data sets ($p = 0.371$) (Table 2).

The mathematical model established by logistic regression was as follows: \( \text{logit}(p = \text{MPN}) = \frac{e^x}{1 + e^x} \), \( x = -2.857 + (1.116 \times \text{mediastinal lymph node enlargement}) + (2.078 \times \text{lobulation}) + (0.997 \times \text{vascular notch sign}) + (1.316 \times \text{spiculation}) + (1.493 \times \text{nodule number}) + \)

### TABLE 1 Characteristics of the participants

| Classification                          | Training data set (n = 363) | Validation data set (n = 129) |
|----------------------------------------|----------------------------|-------------------------------|
|                                        | MPN, N (%) | BPN, N (%) | MPN, N (%) | BPN, N (%) |
| Age                                    |            |            |            |            |
| Mean                                   | 60.1       | 54.1       | 60.8       | 55.4       |
| SD                                     | 11.2       | 12.4       | 11.6       | 13.0       |
| Sex                                     |            |            |            |            |
| Male                                    | 175 (64.6) | 54 (58.7)  | 45 (48.9)  | 20 (54.1)  |
| Female                                  | 96 (35.4)  | 38 (41.3)  | 47 (51.1)  | 17 (45.9)  |
| Smoker                                  |            |            |            |            |
| Yes                                     | 122 (45.0) | 26 (28.3)  | 27 (29.3)  | 13 (35.1)  |
| No                                      | 149 (55.0) | 66 (71.7)  | 65 (70.7)  | 24 (64.9)  |
| Drinker                                 |            |            |            |            |
| Yes                                     | 59 (21.8)  | 19 (20.7)  | 12 (13.0)  | 7 (18.9)   |
| No                                      | 212 (78.2) | 73 (79.3)  | 80 (87.0)  | 30 (81.1)  |
| Family cancer history                   |            |            |            |            |
| Yes                                     | 25 (9.2)   | 3 (3.3)    | 10 (10.9)  | 5 (13.5)   |
| No                                      | 246 (90.8) | 89 (96.7)  | 82 (89.1)  | 32 (86.5)  |
| CEA (ng/ml)                             |            |            |            |            |
| <5                                      | 192 (70.8) | 87 (94.6)  | 51 (55.4)  | 35 (94.6)  |
| ≥5                                      | 79 (29.2)  | 5 (5.4)    | 41 (44.6)  | 2 (5.4)    |
| TNM stage                               |            |            |            |            |
| I + II                                  | 85 (31.4)  | 22 (24.0)  |            |            |
| III + IV                                | 175 (64.6) | 58 (63.0)  |            |            |
| Uncertain                               | 11 (4.0)   | 12 (13.0)  |            |            |
| Pathological type                       |            |            |            |            |
| SCLC                                    | 30 (11.1)  | 11 (12.0)  |            |            |
| LUSC                                    | 67 (24.7)  | 15 (16.3)  |            |            |
| LUAD                                    | 156 (57.6) | 61 (66.3)  |            |            |
| Chronic inflammation                    |            |            |            |            |
| Pulmonary tuberculosis                  |            |            |            |            |
| Granuloma                               |            |            |            |            |
| Others                                  | 18 (6.6)   | 36 (39.1)  | 5 (5.4)    | 6 (16.2)   |

Abbreviations: BPN, benign pulmonary nodule; CEA, carcinoembryonic antigen; LUAD, lung adenocarcinoma; LUSC, squamous cell lung carcinoma; MPN, malignant pulmonary nodule; SCLC, small cell lung cancer.
where $e$ is the natural logarithm, and mediastinal lymph enlargement (0: no, 1: yes), lobulation (0: no, 1: yes), vascular notch sign (0: no, 1: yes), spiculation (0: no, 1: yes), and nodule number (0: >1, 1: 1) are derived from the CT report.

### 3.4 Model comparison

Compared with the CT and lnCEA alone, the model had higher AUC (95%), sensitivity, and specificity, which made up for the deficiency and combined the advantages of the two detection methods. Not significantly, but it reduces the FPR of CT (25.6% – 17.8%) (Table 3).

Based on the validation data set, the logistic model displayed a higher AUC of 0.882 (Figure 4 and Table 4) than the Peking University model; which formula is logit $(p = MPN) = e^x/(1 + e^x)$,

$$x = -4.496 + (0.07 \times \text{age}) + (0.676 \times \text{diameter}) + (0.736 \times \text{spiculation}) + (1.267 \times \text{family history of cancer}) - (1.615 \times \text{calcification}) - (1.408 \times \text{border});$$

where $e$ is the natural logarithm, and 1 represents yes and 0 represents no in the last four elements (family cancer history, calcification, spiculation, and border), with an AUC of 0.748 (Figure 4 and

### TABLE 2 The diagnostic efficiency of the six models in training and validation data sets

| Model   | Training data set | Validation data set | Delong test |
|---------|-------------------|---------------------|-------------|
|         | AUC (95% CI)      | Se (%) | Sp (%) | Accuracy (%) | AUC (95% CI) | Se (%) | Sp (%) | Accuracy (%) | P value |
| Logistic | 0.912 (0.881, 0.944) | 84.1  | 84.8  | 84.3        | 0.882 (0.824, 0.940) | 80.4  | 75.7  | 79.1        | 0.371   |
| RBF     | 0.788 (0.741, 0.835) | 68.6  | 82.6  | 72.1        | 0.831 (0.763, 0.899) | 71.7  | 94.6  | 78.3        | 0.309   |
| MLP     | 0.783 (0.737, 0.829) | 64.2  | 87.0  | 70.0        | 0.825 (0.756, 0.894) | 68.5  | 89.2  | 74.4        | 0.327   |
| C5.0    | 0.751 (0.698, 0.805) | 68.6  | 88.0  | 73.5        | 0.741 (0.652, 0.830) | 59.8  | 89.2  | 68.2        | 0.848   |
| CHAID   | 0.719 (0.664, 0.775) | 51.3  | 95.7  | 62.6        | 0.666 (0.562, 0.770) | 33.7  | 94.6  | 51.2        | 0.374   |
| SVM     | 0.778 (0.727, 0.830) | 68.6  | 80.4  | 71.6        | 0.759 (0.675, 0.843) | 64.1  | 81.1  | 69.0        | 0.701   |

Abbreviations: AUC, area under the curve; CHAID, chi-squared automatic interaction detection; MLP, multilayer perceptron; RBF, radial basis function; Se, sensitivity; Sp, specificity; SVM, support vector machines.
Table 4). Meanwhile, the logistic model showed higher AUC than the Mayo model \((AUC = 0.745)\) (Figure 4 and Table 4); the formula is logit \(p = \frac{MPN}{1 + e^x} = \frac{1}{1 + e^{-6.8272 + (0.0391 \times \text{age}) + (0.7917 \times \text{smoking history}) + (1.3388 \times \text{cancer history}) + (0.1274 \times \text{diameter}) + (1.0407 \times \text{spiculation}) + (0.7838 \times \text{the upper lobe})}\), \(e\) is the natural logarithm, and 1 represents yes and 0 represents no in the four elements (smoking history, cancer history, spiculation, and the upper lobe).

### 3.5 Model visualization

A nomogram was produced according to the logistic regression model, different variables were assigned different scores, and finally the total score was obtained corresponding to the probability of patients with MPN via R. As shown in Figure 5, we selected a BPN patient as sample, and the values of the CT variables and lnCEA were as follows: vascular notch sign = 0, mediastinal lymph node enlargement = 0, mediastinal lymph node enlargement = 0, spiculation = 0, nodule number = 0, and loculation = 0; lnCEA = −0.2. Finally, we obtained the probability of MPN as 4.75%.

### 3.6 Stratified analyses

In the CEA-negative PNs, as shown in Table S2, the logistic model showed an AUC (95% CI) of 0.891 (0.851, 0.932) with 78.6% sensitivity, 87.4% specificity, and 93.2% positive predictive value (PPV).
In the intermediate-size (10-30 mm) PNs, the logistic model indicated an AUC (95% CI) of 0.856 (0.797, 0.916) with 77.6% sensitivity, 78.9% specificity, and 88.3% PPV. For the solitary PNs, the logistic model displayed an AUC (95% CI) of 0.888 (0.842, 0.934) with 86.6% sensitivity, 77.1% specificity, and 94.3% PPV. Meanwhile, in the solid PNs (SPNs), the logistic model showed an AUC (95% CI) of 0.928 (0.898, 0.960) with 88.5% sensitivity, 83.3% specificity, and 94.3% PPV. In addition, the logistic model exhibited an AUC (95% CI) of 0.857 (0.802, 0.911) for early-stage (TNM I-II) MPNs with 70.6% sensitivity, 83.3% specificity, and 94.3% PPV; and 0.912 (0.879, 0.945) for non-small cell lung cancer (NSCLC) with 81.1% sensitivity, 84.8% specificity, and 93.5% PPV.

3.7 | Intermediate-size nodule analysis

In the intermediate-size (10-30 mm) nodule category, proportion of cases between MPNs and BPNs was similar (Figure S1A,B). As for the early-stage lung cancer, the logistic model showed an AUC (95% CI) of 0.822 (0.743, 0.900) with 67.3% sensitivity, 78.9% specificity, 77.1% PPV (Figure 6A, Table 5). The CEA-negative, solid nodules had a high proportion in the category (Table 5). In the CEA-negative PNs, the logistic model displayed an AUC (95% CI) of 0.835 (0.764, 0.905) with 72.3% sensitivity, 83.3% specificity, and 88.2% PPV (Figure 6B and Table 5). For the SPNs, the logistic model exhibited an AUC (95% CI) of 0.885 (0.824, 0.946) with 85.1% sensitivity, 78.1% specificity, and 87.5% PPV (Figure 6C and Table 5).

4 | DISCUSSION

Due to its noninvasiveness and high sensitivity, CT is widely applied in screening high-risk lung cancer. In the present study, 14 CT characteristics were collected, and 5 CT characteristics (nodule number, mediastinal lymph node enlargement, vascular notch sign, spiculation, and lobulation) were selected as independent risk factors for differentiating MPNs from BPNs through univariate, ROC, and multivariable analysis.

Numerous studies have shown that the lobulation can effectively distinguish BPNs and MPNs. Similarly, lobulation as an independent CT predictor for distinguishing MPNs from BPNs had the highest AUC of 0.783 in the study, followed by spiculation (AUC = 0.750, p < 0.001) and vascular notch sign (AUC = 0.675, p < 0.001). As reported, spiculation, a numerous, radial, short, and straight-line shadow of the nodule, is a specific sign of lung cancer. Vascular notch sign, one or several vessels around the PNs enlarged and connected with the nodule, has been reported to participate in distinguishing MPNs from BPNs. Previous research confirmed that multiple nodules tend to be inflammatory nodules or metastatic lung cancer. A solitary PN represents a common diagnostic challenge in clinical practice. In our study, nodule number exhibited an excellent capability for distinguishing MPNs from BPNs (AUC = 0.625, p < 0.001). Lymph node enlargement was related to inflammatory stimulation and tumor metastasis normally, which were identified to associate with the metastasis of lung cancer. In the study, mediastinal lymph node enlargement could distinguish BPNs from MPNs effectively (AUC = 0.657, p < 0.001).

Though CT including low-dose CT (LDCT) could detect small nodules effectively, it had a high FPR of 96.4% in a study on lung cancer screening in a high-risk population. Therefore, it is necessary to combine CT with multiple methods to reduce its FPR and promote its application in clinical practice.
been reported to play an important role in distinguishing MPNs from BPNS. Also, serum biomarkers have the advantage of being noninvasive, convenient to access, and low cost, which prompted its extensive application.

The abnormal expression of serum CEA assisted in the diagnostic of lung cancer in the clinic. For distinguishing MPNs from BPNS, serum CEA has a higher specificity than other serum biomarkers, such as neuron-specific enolase (NSE), with low sensitivity. In this study, we converted CEA to lnCEA via the transformation algorithm to narrow the data range, eliminate heteroscedasticity, and make the data stable.

Data mining is a new technology that emerged with the development of database and artificial intelligence technology as a further extension of the statistical method. It could help researchers to mine and discover latent information and rules from data. Currently, the commonly used methods are logistic regression, ANN, SVM, and DT. Logistic regression model is frequently used in data mining to support disease diagnosis, including distinguishing MPN from BPN. The ANN model is a simplified intelligent model constructed by mathematical methods from the perspective of information processing by imitating the signal transmission between human brain neurons, which can be used for classification and prediction. SVM is based on structural risk minimization, which has great advantages in building high-risk models with small sample sizes. DT is a classification method that follows a tree structure and is the earliest and most important method used to solve classification problems. It can effectively solve the problems caused by linearity and nonlinearity between variables. In the present study, six diagnostic models based on four different algorithms were constructed via SPSS Modeler. Through comparison, the logistic model had excellent diagnostic efficiency compared with the others. The logistic model displayed an AUC of 0.912 in the training data set with 84.1% sensitivity and 84.4% specificity. The model is robust in the validation data set with an AUC of 0.882.

When combined the two data sets, the diagnostic FPR of the logistic model reduced 7.8% (25.6% – 17.8%) compared with the separate application of CT (Table 3). At the same time, compared with lnCEA, with the clinical criteria (CEA = 5 ng/ml) as the cut-off value, the logistic model improved the diagnostic sensitivity from 33.1% to 83.2%, with the specificity having decreased only by 12.4%. The above results showed that the logistic model could combine the advantages of the CT and CEA to distinguish MPNs from BPNS effectively and made up for the disadvantages of their application alone.

In stratified analyses, the logistic model displayed a stable and favorable ability to differentiate MPNs and BPNS. Especially, for CEA-negative patients, the logistic model showed 93.2% PPV and 81.4% accuracy, thus improving the sensitivity of CEA. For PN with a diameter <10 mm, the logistic model exhibited good performance with a specificity of 100.0%. As for early-stage MPNs and BPNS, the logistic model displayed a benign PPV of 88.1%.

**Figure 5** Nomogram for predicting the probability of MPN. As an example, Add the scores achieved for each CT characteristic and lnCEA, obtained the sum on the "total points" axis. Draw a line straight down to determine the probability of MPN. **p < 0.01, ***p < 0.001. MLNE, mediastinal lymph node enlargement; MPN, malignant pulmonary nodule; NO, nodule number; Pr(MPN), probability of MPN; VNS, vascular notch sign.**
Moreover, the logistic model could distinguish MPNs with different pathological types (small cell lung cancer, NSCLC, lung squamous cell carcinoma, and lung adenocarcinoma) from BPNs accurately, with an AUC range from 80.1% to 95.9% at 84.8% specificity.

Guidelines for PNs proposed different monitoring schemes based on the diameter. Among the PNs, intermediate nodules (10-30 mm) occupied a large proportion, which was difficult to identify. Patz reported that the CART model combining nodule size, α₁-antitrypsin (AAT), CEA, and SCC had a PPV of 88% in distinguishing MPNs from BPNs with intermediate diameter. In the present study, the logistic model displayed an excellent differential diagnostic ability for MPNs and BPNs in the intermediate diameter subgroup, where the AUC was 0.856 with 88.3% PPV. Besides, for CEA-negative PNs and solid PNs, the logistic model displayed AUCs of 0.835 and 0.885, respectively.

The Peking University model and Mayo model were the earlier lung cancer prediction models for differentiating MPNs and BPNs. The indicators of the Peking University model and Mayo model were the clinical and radiological variables; serum biomarkers were not involved. Univariate and multivariate analyses were applied to screen independent risk factors, and spiculation was selected as significant indicator in distinguishing MPNs from BPNs in the three models (Peking University model, Mayo model, and the logistic model). Calcification, diameter, and border which were independent risk factors in the Peking University or Mayo models showed no significant difference in distinguishing MPNs from BPNs in our study, that might due to differences in the inclusion criteria or sample heterogeneity. The logistic model possessed higher diagnostic value (AUC = 0.882) than the Peking University model (AUC = 0.712) and the Mayo model (AUC = 0.745) based on the validation data set.

The study also has its limitations. Given the complexity of tumorigenesis, a single biomarker could easily introduce bias to the predictive model. Additional traditional biomarkers and novel biomarkers should be merged, and more efficient strategies should be developed to improve the clinical management of PNs. Aim to improve the diagnostic efficiency, fusion algorithms need to be explored. Also, this study is a retrospective single-center study with limited participants.

**CONCLUSION**

The logistic model combined five CT characteristics, and lnCEA had a favorable diagnostic efficiency in distinguishing MPNs from BPNs. Moreover, with regard to distinguishing CEA-negative and intermediate-diameter MPNs from BPNs in the early stage, the logistic model had an outstanding performance. In the follow-up research, the sample size will be further expanded and multicenter, prospective studies will be performed to verify the stability of the model and make contributions to the distinguishing of lung cancer.

**ACKNOWLEDGMENTS**

We gratefully acknowledge Henan Key Laboratory of Pharmacology for Liver Diseases for providing platform support.
FUNDING INFORMATION
This work was supported by the Leading Talents of Science and Technology Innovation in Henan Province (Grant Number 20420051008), the National Natural Science Foundation of China (Grant Number 8167291), the Key Project of Discipline Construction of Zhengzhou University (Grant Number XKZDQY202009), and the Project of Basic Research Fund of Henan Institute of Medical and Pharmacological Sciences (Grant Number 2022BP0104).

DISCLOSURE
The authors have no conflict of interest.

ETHICAL APPROVAL
Approval of the research protocol: The research protocol was approved by the Medical Ethics Committee of the First Affiliated Hospital of Zhengzhou University (2021-KY-1057-002).

Informed Consent: NA.

Animal Studies: N/A.

REFERENCES
1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209-249.
2. Society AC. Cancer Facts & Figures 2020. American Cancer Society; 2020.
3. Schwartz AG, Cote ML. Epidemiology of lung cancer. Adv Exp Med Biol. 2016;893:21-41.
4. Schabath MB, Cote ML. Cancer Progress and priorities: lung cancer. Cancer Epidemiol Biomarkers Prev. 2019;28:1563-1579.
5. Dickson JL, Horst C, Nair A, Tisi S, Prendercki R, Janes SM. Hesitancy around low-dose CT screening for lung cancer. Ann Oncol. 2022;33:34-41.
6. National Lung Screening Trial Research T, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395-409.
7. Kowall B, Jockel KH, Stang A. Lungenkrebscreening: Aktuelle Entwicklungen [Lung cancer screening: current trends]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2018;61:1551-1558.
8. Sagawa M. Current status and prospects of lung cancer screening. Kyobu Geka. 2021;74:74-83.
9. Gold P, Freedman SO. Demonstration of tumor-specific antigens in human colonic Carcinomata by immunological tolerance and absorption techniques. J Exp Med. 1965;121:439-462.
10. Martin Mateo MC, Rabadán J, Boustamante J. Comparative analysis of certain metals and tumor markers in bronchopulmonary cancer and colorectal cancers. Metals and tumor markers in the neoplastic process. Clin Physiol Biochem. 1990;8:261-266.
11. Zheng J, Ye X, Liu Y, Zhao Y, He M, Xiao H. The combination of CTCs and CEA can help guide the management of patients with SPNs suspected of being lung cancer. BMC Cancer. 2020;20:106.
12. Wu LX, Li XF, Chen HF, et al. Combined detection of CEA and CA125 for the diagnosis for lung cancer: a meta-analysis. Cell Mol Biol. 2018;64:67-70.
13. Grunnet M, Sorensen JB. Carcinoembryonic antigen (CEA) as tumor marker in lung cancer. Lung Cancer. 2012;76:138-143.
14. Xu CH, Xue JS, Zhang XW, Yu LK, Lin Y. The value of macrophage inhibitory cytokine-1 level in differentiating benign from malignant solitary pulmonary nodules. Clin Respir J. 2018;12:1473-1478.
15. Lai Y, Wang X, Zeng T, et al. Serum VEGF levels in the early diagnosis and severity assessment of non-small cell lung cancer. J Cancer. 2018;9:1538-1547.
16. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Arch Intern Med. 1997;157:849-855.
17. Li Y, Chen KZ, Wang J. Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary pulmonary nodules in Chinese people. Clin Lung Cancer. 2011;12:313-319.
18. Qi LL, Wu BT, Tang W, et al. Long-term follow-up of persistent pulmonary pure ground-glass nodules with deep learning-assisted nodule segmentation. Eur Radiol. 2020;30:744-755.
19. Chu ZG, Li WJ, Fu BJ, Lv FJ. CT characteristics for predicting invasiveness in pulmonary pure ground-glass nodules. AJR Am J Roentgenol. 2020;215:351-358.
20. Liu Y, Chen M, Guo C, et al. Clinical-radiological-pathological characteristics of 297 cases of surgical pathology confirmed benign pulmonary lesions in which malignancy could not be excluded in preoperative assessment: a retrospective cohort analysis in a single Chinese hospital. Zhongguo Fei Ai Za Zhi. 2020;23:792-799.
21. Sun Y, Li C, Jin L, et al. Radiomics for lung adenocarcinoma manifesting as pure ground-glass nodules: invasive prediction. Eur Radiol. 2020;30:3650-3659.
22. Winter A, Aberle DR, Hsu W. External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data. Thorax. 2019;74:551-563.
23. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med. 2013;369:910-919.
24. Bergin C, Castellano RA. Mediastinal lymph node enlargement on CT scans in patients with usual interstitial pneumonitis. AJR Am J Roentgenol. 1990;154:251-254.
25. Osinka K, Zielinska-Krawczyk M, Korczynski P, Gornicka B, Krenke R. Impact of endobronchial ultrasound guided transbronchial needle aspiration on diagnostic yield of bronchoscopy in patients with mediastinal lymph node enlargement. Adv Exp Med Biol. 2016;911:33-43.
26. Liu QX, Zhou D, Han TC, et al. A noninvasive multianalytical approach for lung cancer diagnosis of patients with pulmonary nodules. Adv Sci. 2021;8:2100104.
27. Zhao W, Yu H, Han Z, Gao N, Xue J, Wang Y. Clinical significance of joint detection of serum CEA, SCCA, and bFGF in the diagnosis of lung cancer. Int J Clin Exp Pathol. 2015;8:9506-9511.
28. Zhang ZH, Han YW, Liang H, Wang LM. Prognostic value of serum CYFRA21-1 and CEA for non-small-cell lung cancer. Cancer Med. 2015;4:1633-1638.
29. Wang W, Liu M, Wang J, et al. Analysis of the discriminative methods for diagnosis of benign and malignant solitary pulmonary nodules based on serum markers. Oncol Res Treat. 2014;37:740-746.
30. Moitra D, Mandal RK. Prediction of non-small cell lung cancer histology by a deep ensemble of convolutional and bidirectional recurrent neural network. J Digit Imaging. 2020;33:895-902.
31. Manikandan T, Bharathi N. Lung cancer detection using fuzzy auto-seed cluster means morphological segmentation and SVM classifier. J Med Syst. 2016;40:181.
32. Pak K, Kim K, Kim MH, et al. A decision tree model for predicting mediastinal lymph node metastasis in non-small cell lung cancer with F-18 FDG PET/CT. PLoS One. 2018;13:e0193403.

ORCID
Man Liu https://orcid.org/0000-0002-5448-6144
Liping Dai https://orcid.org/0000-0002-5531-1776
33. Patz EF Jr, Campa MJ, Gottlin EB, et al. Biomarkers to help guide management of patients with pulmonary nodules. *Am J Respir Crit Care Med*. 2013;188:461-465.

**SUPPORTING INFORMATION**
Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Liu M, Zhou Z, Liu F, et al. CT and CEA-based machine learning model for predicting malignant pulmonary nodules. *Cancer Sci*. 2022;113:4363-4373. doi: [10.1111/cas.15561](https://doi.org/10.1111/cas.15561)