International Journal of Chronic Obstructive Pulmonary Disease Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

# The Construction of Primary Screening Model and Discriminant Model for Chronic Obstructive Pulmonary Disease in Northeast China

This article was published in the following Dove Press journal:
*International Journal of Chronic Obstructive Pulmonary Disease*

Xiaomeng Li[1]
Yuhao Guo[2]
Wenyang Li[1]
Wei Wang[1]
Fang Zhang[1]
Shanqun Li[3]

[1]Department of Respiratory and Critical Care Medicine, The First Hospital of China Medical University, Shenyang 110000, People's Republic of China; [2]Department of Mathematics and Statistics, Xi'an JiaoTong University, Xi'an 710049, People's Republic of China; [3]Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200020, People's Republic of China

**Objective:** The diagnosis of chronic obstructive pulmonary disease (COPD) is challenging, especially in the primary institution which lacks spirometer. To reduce the rate of COPD missed diagnoses in Northeast China, which has a higher prevalence of COPD, this study aimed to establish efficient primary screening and discriminant models of COPD in this region.

**Patients and Methods:** Subjects from Northeast China were enrolled from December 2017 to April 2019 from The First Hospital of China Medical University. Pulmonary function tests and questionnaire were given to all participants. Using illness or no illness as the goal for screening models and disease severity as the goal for discriminant models, multivariate linear regression, logical regression, linear discriminant analysis, K-nearest neighbor, decision tree and support vector machine were constructed through R language and Python software. After comparing effectiveness among them, the most optimal primary screening and discriminant models were established.

**Results:** Enrolled were 232 COPD patients (124 GOLD I–II and 108 GOLD III–IV) and 218 normal controls. Eight primary screening models were established. The optimal model was $Y = -1.2562–0.3891X_4$ (education level) $+ 1.7996X_5$ (dyspnea) $+ 0.5102X_6$ (cooking fuel grade) $+ 1.498X_7$ (smoking index) $+ 0.8077X_9$ (family history)$-0.5552X_{11}$ (BMI) $+ 0.538X_{13}$ (cough with sputum) $+ 2.0328X_{14}$ (wheezing) $+ 1.3378X_{16}$ (farmers) $+ 0.8187X_{17}$ (mother's smoking exposure history during pregnancy)$-0.389X_{18}$ (kitchen ventilation) $+ 0.6888X_{19}$ (childhood heating). Six discriminant models were established. The optimal model was decision tree (the optimal variables: dyspnea ($x_5$), cooking fuel grade ($x_6$), second-hand smoking index ($x_8$), BMI ($x_{11}$), cough ($x_{12}$), cough with sputum ($x_{13}$), wheezing ($x_{14}$), farmer ($x_{16}$), kitchen ventilation ($x_{18}$), and childhood heating ($x_{19}$)). The code was established to combine the discriminant model with computer technology.

**Conclusion:** Many factors were related to COPD in Northeast China. Stepwise logistic regression and decision tree were the optimal screening and discriminant models for COPD in this region.

**Keywords:** chronic obstructive pulmonary disease, screening, discriminant, severity, model

## Introduction

Chronic obstructive pulmonary disease (COPD) is a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation. The incidence of COPD is particularly high in developing countries. In 2015, the number of COPD patients in China was nearly 100 million.[1] In 2017, a large-scale prospective study conducted by Zhou et al[2] confirmed that early

Correspondence: Wei Wang
Email wwbycmu@126.com



submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/COPD.S250199

intervention and treatment for COPD significantly slows the decline of lung function in patients with early COPD (stage I and II), delays the onset of acute exacerbation, reduces the hospitalization rate and improves the quality of life. These studies make early screening and evaluation for COPD a key issue.

The most accurate and specific tool for early screening, diagnosis and evaluation of COPD is pulmonary function test. However, the high operating cost of the spirometry and its technical requirements for operators make it unavailable for wide use in primary care institutions. It might lead to missed diagnosis and delayed treatment of COPD. Research shows that the all-cause mortality in patients with COPD who were misdiagnosed was 3.1 times as high as in people without airflow limitation, and the risk of contracting pneumonia was increased by 2.7 times in people with COPD. Since spirometry is not recommended in Global Strategy for Diagnosis, Management and Prevention of Chronic Obstructive Lung Disease (GOLD 2019, http://www.goldcopd.com) screening for COPD among asymptomatic individuals, simple, efficient and accurate screening tools for primarily screening COPD at an early stage are important. Some studies have used questionnaires or special spirometric measures in certain patients,[3,4] but the results could not be applied widely because of differences in environment, weather, air pollution and life style among regions. For example, in Southern China, COPD is related to a wet environment and smoking, while in Northern China, cold weather, fuel exposure and closed space during winter contribute to the development of COPD. Thus, a screening tool should consider these factors, especially for primary hospitals.

With the development of information technology, the application of screening models combined with computer technology have become a preferred means of screening for diseases,[5,6] and allowed countless patients to benefit from early diagnosis. Because of the high prevalence of COPD in Northeastern China, most patients with COPD must spend winters in Southern China. This study explored screening and discriminant models specialized for patients with COPD in Northeastern China, especially who live in regions where pulmonary function tests are not available, to help patients be diagnosed and managed as early as possible.

# Patients and Methods

## Participants

The research data were from a database at the outpatient and physical examination center of The First Hospital of China Medical University. From December 2017 to April 2019, 232 patients with COPD were first diagnosed by pulmonologists of The First Hospital of China Medical University and enrolled. GOLD was used to diagnose and categorize COPD severity. Entry criteria were: not diagnosed with COPD before being recruited; age between 40 and 80; airflow limitation ≤70% indicated by forced expiratory volume in one second ($FEV_1$)/forced vital capacity (FVC); and $FEV_1$ reversibility following inhalation of salbutamol <12% of pre-bronchodilator $FEV_1$. Patients with acute exacerbation within the prior 3 months or other respiratory diseases were excluded. Matching by age and gender, we recruited 218 control individuals without respiratory diseases from our hospital. Enrollment criteria for controls were: age between 40 and 80, and no diseases affecting questionnaire filling and lung function tests. All participants were Chinese and we included only permanent residents of Northeast China. All participants were assessed by board-certified pulmonologists. Those with conditions such as mental disease or bronchodilator usage that could influence the results of questionnaires and pulmonary function were excluded. This study was approved by the Ethics Committee of The First Hospital of China Medical University. All participants were informed and agreed to the study.

## Questionnaires

The questionnaire we designed (Table 1) was based on the burden of obstructive lung disease (BOLD) study epidemiological questionnaire, the IPAG-recommended symptom-based COPD questionnaire,[7,8] St. George's respiratory questionnaire,[9] the modified British Medical Research Council questionnaire[10] and COPD assessment test.[11] The questionnaire was adjusted according to GOLD guidelines especially for risk factors in Northeastern China, including demographic data, smoking status, history of fuel exposure, family history, related respiratory symptoms and understanding of the disease. All participants completed the questionnaire on their own or with the assistance of relatives.

## Pulmonary Function Test

Before the test, the safety and accuracy of implementation was evaluated. Participants were required to meet inclusion criteria and take no bronchodilators within 2 weeks. The contra-indications of spirometry testing were as following: had undergone chest, abdomen or eye surgery in the last 3 months; had a heart attack in the last 3 months (eg, angina, myocardial infarction, malignant arrhythmia);

**Table 1** The Questionnaire

| Factor | Question | Option |
|---|---|---|
| $x_1$ | Which is your resident type now? | a. Bungalow<br>b. Building |
| $x_2$ | Gender | a. Male<br>b. Female |
| $x_3$ | Age | Year of birth to 2019 |
| $x_4$ | Which is your education level? | a. Undergraduate and above<br>b. High school<br>c. Junior high school<br>d. Primary school<br>e. Uneducated |
| $x_5$ | Do you have any dyspnea without strenuous exercise recently? | a. Never<br>b. <1 Time/week<br>c. 1–2 Times/week<br>d. 3–6 Times/week<br>e. Daily |
| $x_6$ | Which of the following was the cooking fuel you often use? | a. No<br>b. Electricity<br>c. Natural gas/liquefied gas/biogas<br>d. Coal<br>e. Firewood |
| $x_7$ | How many packs do you smoke per year? | |
| | How many years have you smoked? | |
| $x_8$ | How many days are you exposed to secondhand smoke each year? | |
| | How many years have you been exposed to secondhand smoke? | |
| $x_9$ | Do your parents or siblings have respiratory problems? | a. No<br>b. Yes |
| $x_{10}$ | Have you ever had the following diseases during childhood? (Multiple choice) | a. None<br>b. Pneumonia<br>c. Tuberculosis<br>d. Bronchiectasis |
| $x_{11}$ | BMI ($kg/m^2$) | |
| $x_{12}$ | Will the weather change cause you to cough? | a. Never<br>b. <1 Time/week<br>c. 1–2 Times/week<br>d. >3 Times/week |
| | Do you often cough when you do not have a cold? | a. Never<br>b. <1 Time/week<br>c. 1–2 Times/week<br>d. >3 Times/week |
| $x_{13}$ | Do you often cough with sputum from your chest in the morning? | a. Never<br>b. <1 Time/week<br>c. 1–2 Times/week<br>d. >3 Times/week |
| $x_{14}$ | Do you have any wheezing without strenuous exercise recently? | a. No<br>b. Yes |
| $x_{15}$ | Which of the following is your birth quarter | a. 4–9 months<br>b. 10–3 months |
| $x_{16}$ | Are you a farmer? | a. No<br>b. Yes |
| $x_{17}$ | Has your mother ever been exposed to smoke during pregnancy? | a. Never<br>b. Yes, "She smoked at that time" or "She was exposed to secondhand smoke at that time"<br>c. Yes, "She smoked at that time" and "She was exposed to secondhand smoke at that time" |
| $x_{18}$ | Which of the following is your kitchen ventilation? | a. Range hood<br>b. Ventilation fan<br>c. Chimney<br>d. No |
| $x_{19}$ | What did you use to keep warm in childhood? | a. Central heating<br>b. Electricity or air conditioning<br>c. Coal stove<br>d. Firewood/brazier/fire |
| $x_{20}$ | What do you use for heating at present? | a. Central heating<br>b. Electricity or air conditioning<br>c. Coal stove<br>d. Firewood/brazier/fire |

**Abbreviation:** BMI, body mass index.

hospitalized for heart disease in the last 1 month; massive hemoptysis in the last 1 month; stroke in the last 1 month; receiving anti-TB drug treatment or having active pulmonary tuberculosis; uncontrolled severe hypertension in patients with diastolic pressure greater than 100 mm Hg and a systolic pressure greater than 200 mm Hg; aortic aneurysm; severe hyperthyroidism; medication for seizures; history of retinal detachment; or facial paralysis. Standard pulmonary function instruments (YAEGER, Vmax, Germany) were used. We used a 3-L syringe to calibrate the spirometer daily. Participants were seated,

wearing a nose clip, and using a disposable mouthpiece. Participants were required to have an error of ≤0.15 L between the best value and the next best value of FVC and $FEV_1$ in three acceptable tests. If FVC ≤ 1.0 L, the error was ≤0.10 L. We used the same criteria and administered a bronchodilator (salbutamol 400 μg) via inhalation through a 500-mL spacer and repeated spirometry after 20 min.[12] Criteria for airflow limitation and grading were according to GOLD 2019. Airflow limitation was defined as the fixed ratio of FEV1/FVC <0.70 (post bronchodilator). Severity of airflow limitation was defined as GOLD I (mild, $FEV_1$ ≥80% predicted), GOLD II (moderate, 50% ≤$FEV_1$ <80% predicted), GOLD III (severe, 30% ≤$FEV_1$ <50% predicted) and GOLD IV (very severe, $FEV_1$ <30% predicted).

## Variables

Screening and discriminant variables were determined according to questionnaires (Table 1). They were resident type ($x_1$), gender ($x_2$), age ($x_3$), education level ($x_4$), dyspnea ($x_5$), cooking fuel grade ($x_6$), smoking index ($x_7$), second-hand smoking index ($x_8$), family history ($x_9$), infectious history at child age ($x_{10}$), body mass index (BMI) ($x_{11}$), cough (x12), cough with sputum (x13), wheezing ($x_{14}$), birth quarter ($x_{15}$), farmers ($x_{16}$), maternal pregnancy exposure ($x_{17}$), kitchen ventilation ($x_{18}$), childhood heating ($x_{19}$), and current heating ($x_{20}$). Discriminate factors were also quantitatively assigned. For further standardization, all variables except age, smoking index, second-hand smoking index and BMI were assigned from 0, and the order was based on their influence on the occurrence and development of COPD by GOLD guidelines (Table 2).

## Establishment and Verification of Optimal Primary Screening and Discriminant Models

Before building the model, we completed missing data from the questionnaire collection process. Missing data completion methods were: variables 3, 7, 8, and 11 were completed using the mean method; variables 1, 2, and 3 were completed using the mode method; and the remaining variables were completed using the median method. Second, we performed Z-score standardization on the data set, X` = (X-mean)/standard deviation, and converted the corresponding variables to a distribution with mean 0 and variance 1 to eliminate the influence of dimension. Regularizing operations on different models were performed to eliminate the effect of overfitting the model on the prediction results. COPD primary screening and discriminant models were constructed using general linear regression (multivariate linear regression), generalized linear regression (logistic regression), linear discriminant analysis, K-nearest neighbor, decision tree, conditional decision tree and support vector machine method. Two hundred and thirty-two COPD patients and 218 control groups were randomly selected. The data set was split 4:1 by stratified random sampling and four-fifths was used as a training group to establish models (360, training set). One-fifth was used as a test group to test models (90, test set). Due to the resampling methods, bootstrapping or cross-validation was more powerful than splitting the sample for internal validation.[13] We applied cross-validation on the basis of random stratification. By comparing $F_1$ value, accuracy, recall rate, area under curve (AUC) value and precision, the optimal primary screening model was chosen. Multicollinearity was calculated to assess the feasibility of the optimal model. Receiver operating characteristic (ROC) curves and confusion matrix were constructed to describe the screening effectiveness of the optimal model. By analyzing total accuracy, the optimal discriminant model was chosen. The establishment and verification of the optimal model was:

(1) Primary screening model

We used l2 regularization to constrain the objective function in the optimization process of logistic regression to eliminate the influence of overfitting the model on the prediction results and improve the generalization ability of the model. (Regularization parameter C = 1)

$$\min_{\beta,c} \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n}\log\left(\exp(-y_i(X_i\beta + c) + 1)\right)$$

Accord$_i$ng to a given set of patient samples $T=\{x_{1i},x_{2i}, \ldots \ldots, y_{2i}\}^n_{i=1}$, $x_{1i},x_{2i} \ldots \ldots$ was a series of characteristic attributes of the i-th patient and $y_{2i} \in \{0,1\}$ was a two-category attribute variable ($y_{2i}$ = 0 indicated that the i-th patient did not have COPD, $y_{2i}$=1 indicated that the i-th patient had COPD). The optimal COPD primary screening model was established by stepwise logistic regression.

According to the logic function (1)

$$p_i = \frac{e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots\ldots)}}{1+e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots\ldots)}} \quad (1)$$

**Table 2** The Variable Assignment

| Factor | Variable | Quantification |
|---|---|---|
| $x_1$ | Resident type | Building = 0, Bungalow = 1 |
| $x_2$ | Gender | Female=0, Male=1 |
| $x_3$ | Age | Year of birth to 2019 |
| $x_4$ | Education level | Undergraduate and above=0, High school=1, Junior high school=2, Primary school=3, Uneducated=4 |
| $x_5$ | Dyspnea | Never = 0, <1 Time/week = 1, 1–2 Times/week = 2, 3–6 Times/week = 3, Daily = 4 |
| $x_6$ | Cooking fuel grade | No = 0, Electricity = 1, Natural gas/liquefied gas/biogas = 2, Coal = 3, Firewood = 4 |
| $x_7$ | Smoking index | pack*year |
| $x_8$ | Second-hand smoking index | Contact day/year* contact year |
| $x_9$ | Family history | No = 0, Yes =1 |
| $x_{10}$ | Infectious history at child age | None = 0, One type= 1, Two types= 2, Three types= 3 |
| $x_{11}$ | BMI | $kg/m^2$ |
| $x_{12}$ | Cough | None = 0, Level 1 = 1, Level 2 = 2, Level 3 = 3, Level 4 = 4, Level 5 = 5, Level 6 = 6 |
| $x_{13}$ | Cough with sputum | None=0, Level 1=1, Level 2=2, Level 3=3 |
| $x_{14}$ | Wheezing | No = 0, Yes=1 |
| $x_{15}$ | Birth quarter | 4–9 months=0, 10–3 months=1 |
| $x_{16}$ | Farmers | No=0, Yes=1 |
| $x_{17}$ | Mother's smoking exposure history during pregnancy | No = 0, level 1 = 1, level 2 = 2 |
| $x_{18}$ | Kitchen ventilation | Range hood=0, Ventilation fan=1, Chimney=2, No=3 |
| $x_{19}$ | Childhood heating | Central heating = 0, Electricity or air conditioning = 1, Coal stove = 2, Firewood/brazier/fire = 3 |
| $x_{20}$ | Current heating | Central heating=0, Electricity or air conditioning=1, Coal stove=2, Firewood/brazier/fire=3 |
| $y_1$ | $FEV_1$% | Actual value |
| $y_2$ | Binary classification | Normal person: $y_2$=0, COPD patients: $y_2$=1 |
| $y_3$ | Three classification | Normal person: $y_3$=1, Mild/moderate COPD patient: $y_3$=2, Severe/very severe COPD patient: $y_3$=3 |

**Note:** The asterisks (*) in columns $X_7$ and $X_8$ in the table stand for multiplication.
**Abbreviations:** BMI, body mass index; $FEV_1$, forced expiratory volume in one second; COPD, chronic obstructive pulmonary disease.

order:

$$Z_i = \beta_0 + \beta_1\chi_{1i} + \beta_2\chi_{2i} + \ldots \quad (2)$$

the original form became:

$$P_i = \frac{e^{Z_i}}{1 + e^{Z_i}} \quad (3)$$

Here, $P_i = E(y_i = 1|Z_i)$ indicated the probability of a patient having COPD under Zi condition. Hence, $1 - P_i = E(y_i = 0|Z_i)$ presented the probability that the patient did not have COPD under Zi condition.

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \quad (4)$$

To calculate the partial coefficient, we did the regression, which yielded:

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i} \quad (5)$$

$\frac{P_i}{1-P_i}$ was the probability ratio (odds ratio) of a patient with COPD to a patient without COPD. The logarithm gave the equation:

$$L_i = Ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = \beta_0 + \beta_1\chi_{1i} + \beta_2\chi_{2i} + \ldots\ldots \quad (6)$$

This was the logistic regression model reflecting the probability of having COPD in terms of multiple related factors.

Based on the above model, stepwise logistic regression was carried out according to the principle of the lowest Akaike information criterion (AIC) value. The AIC information criterion is a standard to measure the goodness of fit of a statistical model. AIC encouraged the fitting goodness of data and tried to avoid overfitting, so the preferred model had the smallest AIC value.

(2) Discriminant model

According to a given set of patient samples $T=\{x_{1i},x_{2i},\ldots\ldots,y_{3j}\}^{n}_{i=1}$, $x_{1i}$, $x_{2i}$ …… was a series of characteristic attributes of the i-th patient, $y_{3j}\in\{1,3\}$ was a three-category attribute variable ($y_{3j}$ = 1 meant the i-th patient did not have COPD, $y_{3j}$ = 2 meant that the i-th patient had mild/moderate COPD, and $y_{3j}$=3 meant i-th patient had severe/very severe COPD). The optimal discriminant model was established by decision tree.

$$H(T)=\sum_{j-1}^{3}\frac{|y_{3j}|}{|T|}\log_2\frac{|y_{3j}|}{|T|} \quad (7)$$

We calculated the empirical entropy H (T) of the data set T, which indicated the uncertainty of the data set T.

We calculated the empirical conditional entropy H ($T|x_j$) of the feature $x_j$ versus the data set T. The uncertainty of classifying the data set T was given by feature $x_i$.

$$H(T|x_i)=\sum_{k-1}^{x}\frac{|T_k|}{|T|}H(T_k)=\sum_{k-1}^{x}\frac{|T_k|}{|T|}\sum_{j-1}^{3}\frac{|y_{3j}|}{|T|}\log_2\frac{|y_{3j}|}{|T|} \quad (8)$$

We calculated the information gain, which was the reduced degree of uncertainty in the classification of the data set using the feature $x_i$

$$g(T,x_i)=H(T)-H(T|x_i) \quad (9)$$

Therefore, when we chose features for the model, the lower the uncertainty degree, the more the information gain. For the data set T, different features tended to have different information gains, and features with more information gains had stronger classification capabilities. After selecting the optimal feature recursively and dividing the training data according to the feature, the best classification for each subdata set under the current conditions was made. To eliminate the influence of overfitting, we pruned the original decision tree and set the maximum depth to 4 to generate the final decision tree.

## Results

Enrolled were 232 COPD patients aged from 40 to 80 years, 128 males and 104 females. Among them, 124 patients had mild and moderate COPD, and 108 had severe and very severe COPD. In addition, 218 normal subjects aged from 40 to 80 years were enrolled as the control group, 114 males and 104 females. Information about the 232 patients with COPD and 218 control participants is listed in Table 3. Compared to the control group, COPD patients showed some risk factors such as low weight, living in a bungalow, low level of education, and exposure to coal and firewood instead of electricity. After adjusting for gender and age, parameters of birth season, education level, BMI, dyspnea, family history, cough and sputum, wheeze, farmers, resident type, smoking/passive smoking, mother's smoking history during pregnancy, fuel exposure level, childhood heating and current heating had statistical differences between COPD patients and normal controls. Education level, dyspnea, BMI, cooking fuel exposure, cough and sputum, wheeze, farmers, mother's smoking history during pregnancy, kitchen ventilation and current heating were related to the severity of COPD.

## Establishment and Verification of the Primary Screening Model for COPD

Recorded information was used to construct primary screening models. The effectiveness of each model was evaluated to find the optimal screening model for patients with COPD in Northeast China (Tables 4 and 5). Since test set substitution had more practical significance, the stepwise logistic regression prediction model was determined to be the optimal primary screening model.

Logistic regression was performed on the training set with 20 influencing factors ($x_1$-$x_{20}$) used as independent variables. After standardization, corresponding mean and standard deviation were determined and are in Table 6. "Illness or not (y)" was used as a dependent variable for logistic regression, and backward stepwise logistic regression based on AIC values was used to filter the variables. This yielded the equation:

$$Ln\left(\frac{p}{1-p}\right)=-1.2562-0.3891X_4+1.7996X_5+0.5102X_6+1.498X_7+0.8077X_5-0.5552X_{11}+0.538X_{13}+2.0328X_{14}+1.3378X_{16}+0.8187X_{17}-0.389X_{18}+0.6888X_{19}$$

Dependent variables were tested and no multicollinearity ($\sqrt{vif}<2$) was found (Table 7). We calculated variable parameters and significance of the model. The null hypothesis for the regression equation significance test was rejected because the *P* value of some selected variables was less than 0.05 (Table 8), so the relationship between dependent variables was statistically significant. Finally, the model was tested with the test and training sets. The ROC curve of the primary screening model was in Figures 1 and 2. The model had excellent predictability and 0 was the optimal critical point. According to the confusion matrix (Tables 9 and 10), in the training set, sensitivity was 0.9569,

**Table 3** The Comparison of Basic Information Among Three Groups

| Test | COPD | | | Control (n=218) | *P* |
|---|---|---|---|---|---|
| | GOLD I+II (n=124) | GOLD III+IVIV (n=108) | *P* | | |
| **Resident type** | | | 0.02245 | | <0.001 |
| Building | 56 | 33 | | 124 | |
| Bungalow | 68 | 75 | | 94 | |
| **Gender** | | | 0.7097 | | 0.5415 |
| Male (%) | 67 (54%) | 61 (56%) | | 114 (52.2%) | |
| **Age (years)** | | | 0.5525 | | 0.9503 |
| 40–59 | 44 | 37 | | 73 | |
| 60–80 | 76 | 70 | | 144 | |
| Means | 61.30 | 62.03 | | 61.58 | |
| **Education level** | | | <0.001 | | <0.001 |
| Undergraduate and above | 10 | 5 | | 10 | |
| High school | 16 | 7 | | 85 | |
| Junior high school | 58 | 28 | | 98 | |
| Primary school | 26 | 52 | | 23 | |
| Uneducated | 14 | 16 | | 2 | |
| **Dyspnea** | | | <0.001 | | <0.001 |
| Never | 4 | 0 | | 100 | |
| <1 Time/week | 5 | 4 | | 92 | |
| 1–2 Times/week | 33 | 6 | | 23 | |
| 3–6 Times/week | 28 | 23 | | 3 | |
| Daily | 54 | 75 | | 0 | |
| **Cooking fuel grade** | | | <0.001 | | <0.001 |
| No | 0 | 0 | | 0 | |
| Electricity | 6 | 2 | | 96 | |
| Liquefied gas | 58 | 35 | | 92 | |
| Coal | 46 | 40 | | 10 | |
| Firewood | 14 | 31 | | 20 | |
| **Smoking (pack*years)** | 20.1 ±25.8 | 25.6 ±28.8 | 0.1296 | 1.2±3.0 | <0.001 |
| **Second-hand smoking index** | 4703.99 ±5799.85 | 4521.30 ±5576.68 | 0.8081 | 1923.62 ±3078.06 | <0.001 |
| **Family history** | | | 0.07935 | | <0.001 |
| No | 66 | 45 | | 169 | |
| Yes | 58 | 63 | | 49 | |
| **Infectious history at child age** | | | 0.06133 | | 0.7544 |
| None | 87 | 61 | | 142 | |
| One | 22 | 28 | | 46 | |
| Two | 11 | 14 | | 22 | |
| Three | 4 | 5 | | 8 | |
| **BMI (kg/m$^2$)** | 23±4 | 21.50 ±3.48 | <0.001 | 25±3 | <0.001 |
| **Cough** | | | <0.001 | | <0.001 |
| None | 7 | 3 | | 14 | |
| Level 1 | 3 | 4 | | 2 | |
| Level 2 | 16 | 4 | | 143 | |
| Level 3 | 18 | 5 | | 47 | |
| Level 4 | 25 | 21 | | 11 | |
| Level 5 | 21 | 22 | | 0 | |
| Level 6 | 34 | 49 | | 1 | |
| **Cough with sputum** | | | 0.00607 | | <0.001 |
| None | 12 | 9 | | 16 | |
| Level 1 | 23 | 9 | | 166 | |
| Level 2 | 40 | 23 | | 35 | |
| Level 3 | 49 | 67 | | 1 | |
| **Wheeze** | | | 0.06013 | | <0.001 |
| No | 4 | 0 | | 188 | |
| Yes | 120 | 108 | | 30 | |
| **Birth quarter** | | | 0.7445 | | <0.001 |
| First and fourth quarters (%) | 72 (58%) | 65 (60%) | | 87 (40%) | |
| **Farmers** | | | <0.001 | | <0.001 |
| No | 63 | 30 | | 205 | |
| Yes | 61 | 78 | | 13 | |
| **Mother's smoking exposure history during pregnancy** | | | 0.0083 | | <0.001 |
| None | 53 | 30 | | 168 | |
| Level 1 | 44 | 41 | | 36 | |
| Level 2 | 27 | 37 | | 14 | |
| **Kitchen ventilation** | | | 0.00628 | | <0.001 |
| Range hood | 59 | 28 | | 177 | |
| Ventilation fan | 7 | 2 | | 13 | |
| Chimney | 36 | 66 | | 22 | |
| No | 22 | 12 | | 6 | |

(*Continued*)

**Table 3** (Continued).

(*Continued*)

**Table 3** (Continued).

| Test | COPD | | | Control (n=218) | P |
|---|---|---|---|---|---|
| | GOLD I+II (n=124) | GOLD III+IVIV (n=108) | P | | |
| **Heating history during Childhood** | | | 0.2376 | | <0.001 |
| Centralized heating | 9 | 1 | | 116 | |
| Electricity | 0 | 1 | | 14 | |
| Coal | 50 | 50 | | 69 | |
| Firewood | 65 | 56 | | 19 | |
| **Current heating** | | | <0.001 | | <0.001 |
| Centralized heating | 68 | 31 | | 194 | |
| Electricity | 1 | 0 | | 1 | |
| Coal | 29 | 39 | | 21 | |
| Firewood | 26 | 38 | | 2 | |

**Note:** Data were given as n or mean±SD.
**Abbreviation:** BMI, body mass index.

specificity was 0.948, positive predictive value was 0.951, and negative predictive value was 0.953. In the test set, sensitivity was 0.956, specificity was 0.977, positive predictive value was 0.978and negative predictive value was 0.956. In the test set, accuracy was 0.9667, F1 value was 0.9670and AUC was 0.967.

## Establishment and Verification of the COPD Discriminant Model

According to Table 11, the decision tree model (test set accuracy 0.8333, training set with cross-validation accuracy 0.8361) was the optimal discriminant model because the results of the test set had more clinical significance and research value. Thus, we used Python software to establish a decision tree model for the training set (Figure 3) and tested our discriminant model. Since the branch of the tree was not complicated, we decided not to prune the model in order to protect the amount of variables (Figure 3). The information value of the COPD discriminant model is in Table 9. We chose 10 for Nsplit because the error no longer changed at this level. The parameter trend graph of COPD discriminant model is in Figure 3. The confusion matrix of the model is in Table 12 (training set with cross-validation accuracy) and Table 13 (test set). In the training set, sensitivity was GOLD I–II 0.737 and GOLD III–IV 0.666, specificity was 0.977, positive predictive value was GOLD I–II 0.73 and GOLD III–IV 0.7, negative predictive value was 0.918, and accuracy was 0.8361. In the test set, sensitivity was GOLD I–II 0.8 and GOLD III–IV 0.619, specificity was 0.95, positive predictive value was GOLD I–II 0.666 and GOLD III–IV 0.866, negative predictive value was 0.933, and accuracy was 0.8333. With computer technology, we turned the optimal discriminant model to the coding program to apply conveniently (Appendix 1).

## Discussion

This study explored high-risk factors for COPD in Northeast China. Several factors were found to be related to development of the disease such as the season of birth, BMI, family history, living environment, mother's smoking history during pregnancy, biofuel exposure and current heating style. Due to the severe cold winter in the Northeast China, especially rural people who live in bungalows usually have a longer wood or coal-burning heating

**Table 4** The Summary of Primary Screening Models for COPD (Training Set with Cross-Validation)

| | Accuracy | Precision | Recall | F1 | AUC |
|---|---|---|---|---|---|
| Multiple linear regression | 0.9278 | 0.9444 | 0.9140 | 0.9290 | 0.928 |
| Stepwise multiple linear regression | 0.9361 | **0.9657** | 0.9086 | 0.9363 | 0.937 |
| Logistic regression | 0.9472 | 0.9563 | 0.9409 | 0.9485 | 0.947 |
| Stepwise logistic regression | **0.9528** | 0.9519 | 0.9570 | **0.9544** | **0.953** |
| Linear discriminant analysis | 0.9417 | 0.9188 | **0.9731** | 0.9452 | 0.941 |
| KNN | 0.9361 | **0.9657** | 0.9086 | 0.9363 | 0.937 |
| Decision tree | 0.9278 | 0.9444 | 0.9140 | 0.9290 | 0.928 |
| SVM | 0.9472 | 0.9418 | 0.9570 | 0.9493 | 0.947 |

**Notes:** $F_1$ value, the principal criterion; accuracy, recall rate, AUC value and precision, the secondary criterion; the bold font in the table indicates that they have the highest score in this category. In the comparison of the primary screening models, we should finally select the model with the most points and the highest score based on the F1 value priority. Second, consider the highest accuracy, precision, recall, AUC value.
**Abbreviations:** AUC, area under the curve; KNN, k-nearest neighbor; SVM, support vector machines.

**Table 5** The Summary of Primary Screening Model for COPD (Test Set)

| | Accuracy | Precision | Recall | F1 | AUC |
|---|---|---|---|---|---|
| Multiple linear regression | 0.9333 | 0.9000 | **0.9783** | 0.9375 | 0.932 |
| Stepwise multiple linear regression | 0.9333 | 0.9545 | 0.9130 | 0.9333 | 0.934 |
| Logistic regression | 0.9556 | 0.9565 | 0.9565 | 0.9565 | 0.956 |
| Stepwise logistic regression | **0.9667** | **0.9778** | 0.9565 | **0.9670** | **0.967** |
| Linear discriminant analysis | 0.9333 | 0.9000 | **0.9783** | 0.9375 | 0.932 |
| KNN | 0.9333 | 0.9545 | 0.9130 | 0.9333 | 0.934 |
| Decision tree | 0.9444 | 0.9556 | 0.9348 | 0.9451 | 0.945 |
| SVM | 0.9444 | 0.9184 | **0.9783** | 0.9474 | 0.944 |

**Notes:** $F_1$ value, the principal criterion; accuracy, recall rate, AUC value and precision, the secondary criterion; the bold font in the table indicates that they have the highest score in this category. In the comparison of the primary screening models, we should finally select the model with the most points and the highest score based on the F1 value priority. Second, consider the highest accuracy, precision, recall, AUC value.
**Abbreviations:** AUC, area under the curve; KNN, k-nearest neighbor; SVM, support vector machines.

**Table 6** Corresponding Mean and Standard Deviation of Primary Screening Model Variables After Standardization

| Variables | Mean (μ) | Standard Deviation (σ) |
|---|---|---|
| x4 | 2.02 | 0.99 |
| x5 | 2.01 | 1.56 |
| x6 | 2.27 | 0.97 |
| x7 | 12.27 | 22.49 |
| x9 | – | – |
| x11 | 23.54 | 3.55 |
| x13 | 1.66 | 0.95 |
| x14 | – | – |
| x16 | – | – |
| x17 | 0.62 | 0.76 |
| x18 | 0.87 | 1.10 |
| x19 | 1.72 | 1.18 |

**Table 8** The Variable Parameters and Significance of Primary Screening Model for COPD

| | Estimate | Std. | Error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|---|
| (Intercept) | −1.2562 | 2.7190 | −1.679 | 0.0931 | . |
| x4 | −0.3891 | 0.3162 | −2.236 | 0.0254 | * |
| x5 | 1.7996 | 0.3372 | 3.961 | 0.0001 | *** |
| x6 | 0.5102 | 0.4164 | 2.026 | 0.0428 | * |
| x7 | 1.4980 | 0.0406 | 3.088 | 0.0020 | ** |
| x9 | 0.8077 | 0.6883 | 2.300 | 0.0214 | * |
| x11 | −0.5552 | 0.0912 | −2.173 | 0.0298 | * |
| x13 | 0.5380 | 0.3372 | 1.931 | 0.0535 | . |
| x14 | 2.0328 | 0.9837 | 3.246 | 0.0012 | ** |
| x16 | 1.3378 | 0.7777 | 2.839 | 0.0045 | ** |
| x17 | 0.8187 | 0.5023 | 3.125 | 0.0018 | ** |
| x18 | −0.3890 | 0.3177 | −2.293 | 0.0218 | * |
| x19 | 0.6888 | 0.3094 | 1.934 | 0.0532 | . |

**Note:** The *p* is the most significant between 0 and 0.001, indicated by '***'; The *p* is extremely significant between 0.001 and 0.01, indicated by '**'; The *p* is relatively significant between 0.01 and 0.05, indicated by '*'; The *p* is significant between 0.05 and 0.1, indicated by '.'.

time, and a higher chance for exposure to biofuel. This study found that the above factors were closely related to occurrence and severity of COPD and were different from factors in the southern area of China. In Northeast China, among the primary screening and discriminant models constructed with data on high-risk factors of COPD patients, a logistic regression model was the most effective for primary screening and a decision tree model was the best for discrimination. Combined with the computer technology, both models could be applied conveniently and accurately for COPD assessment by inputting the related factors. This study investigated the influence of regional characteristics of Northeast China on patients with COPD. An optimal primary screening model and a discrimination model were established by statistically comparing different models that were particularly suitable for primary hospitals.

Besides age, gender, education and smoking history, which were known as risk factors, some special factors were related to the occurrence of COPD in Northeast

**Table 7** Multicollinearity Analysis of Independent Variables

| Variable | Vif Value | Variable | Vif Value | Variable | Vif Value | Variable | Vif Value |
|---|---|---|---|---|---|---|---|
| x4 | 1.57 | x7 | 1.40 | x13 | 1.18 | x17 | 1.47 |
| x5 | 2.01 | x9 | 1.39 | x14 | 1.85 | x18 | 1.84 |
| x6 | 1.69 | x11 | 1.23 | x16 | 1.51 | x19 | 1.30 |

**Abbreviation:** Vif value, variance inflation factor value.

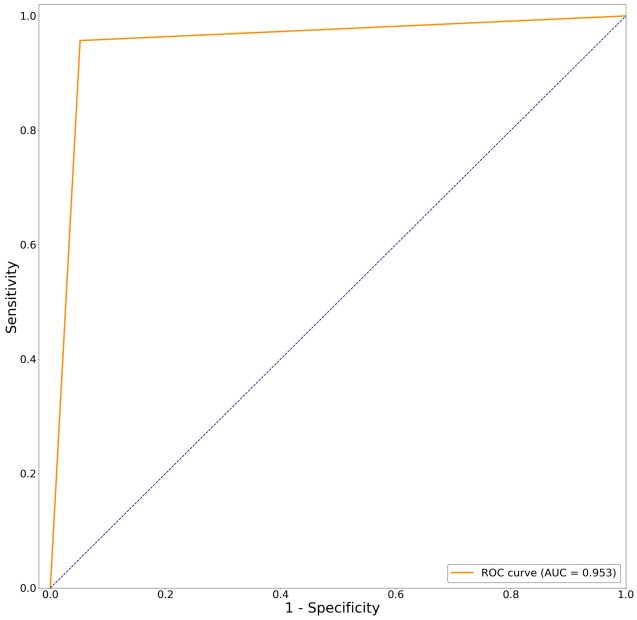


**Figure 1** The ROC curve of the primary screening model (training set with cross-validation).

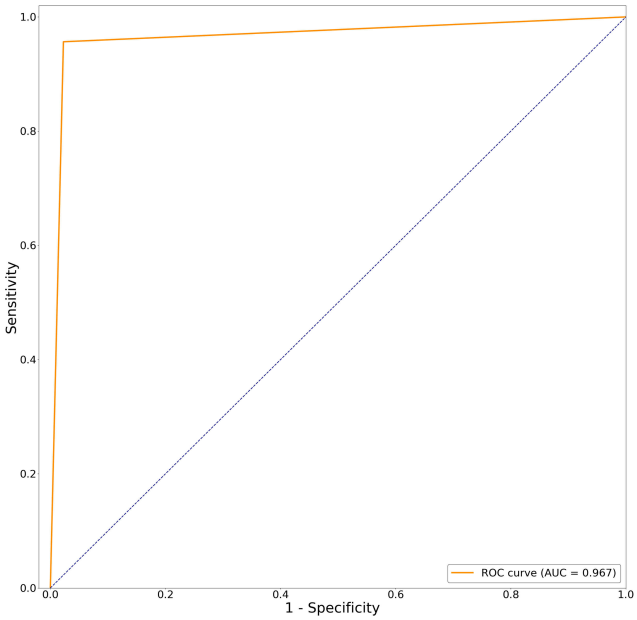


**Figure 2** The ROC curve of the primary screening model (test set).

China such as family history, mother's smoking history during pregnancy, birth season, resident type, BMI, fuel exposure, kitchen ventilation and heating style. The last four factors also contributed to COPD severity. First, our research found that family history was an important factor for predicting the occurrence of COPD, consistent with a study by McCloskey et al.[14] Although no studies documented hereditary deficiency of alpha-1 antitrypsin

**Table 9** The Confusion Matrix of Primary Screening Mode l (Training Set with Cross-Validation)

| Truth \ Prediction | y2=0 | y2=1 |
|---|---|---|
| y2=0 | 165 | 9 |
| y2=1 | 8 | 178 |

**Table 10** The Confusion Matrix of Primary Screening Model (Test Set)

| Truth \ Prediction | y2=0 | y2=1 |
|---|---|---|
| y2=0 | 43 | 1 |
| y2=1 | 2 | 44 |

**Table 11** The Effectiveness of Different Discriminant Models

| | Accuracy (Training Set with Cross-Validation) | Accuracy (Test Set) |
|---|---|---|
| Multiple linear regression | 0.7778 | 0.7667 |
| Stepwise multiple linear regression | 0.7944 | 0.7889 |
| Linear discriminant analysis | 0.8028 | 0.7889 |
| KNN | 0.8139 | 0.8000 |
| Decision tree | **0.8361** | **0.8333** |
| SVM | 0.8278 | 0.8222 |

**Notes:** The bold font in the table indicates that they have the highest score in this category. In the comparison of discriminant models, we have to select the model with the highest accuracy value.
**Abbreviations:** KNN, k-nearest neighbor; SVM, support vector machines.

(AATD), we hypothesize that Asians may be affected by certain genes since a change in the gene encoding matrix metalloproteinase 12 (MMP12) is reported to be related to COPD in Asians. Chinese scholars[15] reported that the glutathione S-transferase M1(GSTM1) null, GSTT1 null, and combined GSTM1/glutathione S-transferase theta 1(GSTT1) null genotypes might be risk factors for development of COPD. The GSTT1 null polymorphism showed association with only Asian COPD patients. Thus, genetics with environmental factors may influence the susceptibility to disease among specific populations.[14] The exact factors that led to "familial aggregation" in Asia such as similar living environments and lifestyle or some potential genes, deserved to be further investigated. Second, Tager et al[16] found that smoking during pregnancy-imposed risks on the fetus and affected the development of the lungs and immune system during the first 18 months

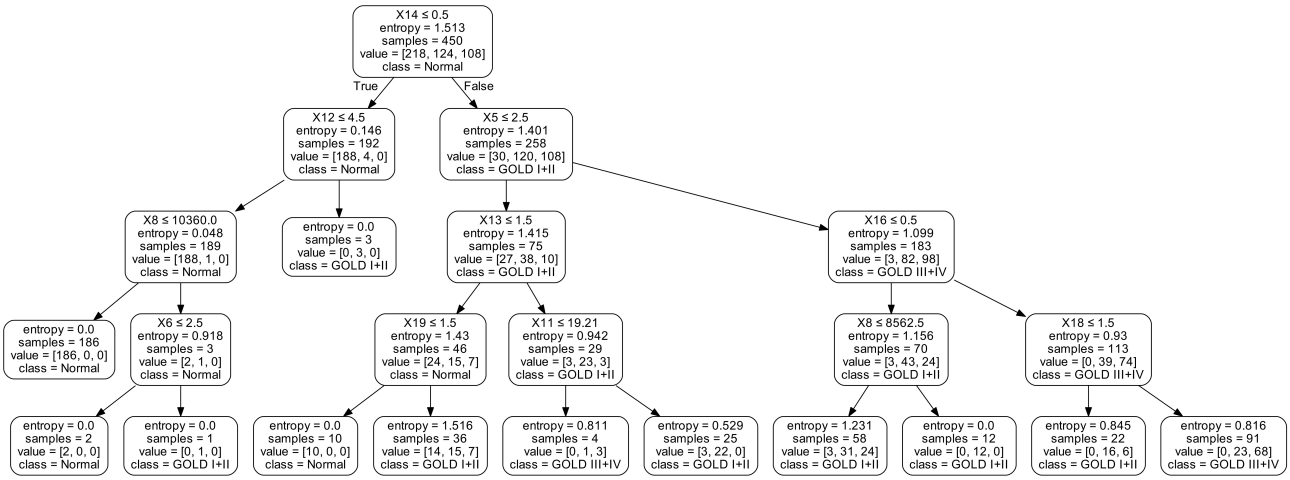


**Figure 3** The discriminant model for COPD.

of life. Some studies further observed that exposure to smoking during childhood and adolescence affects lung growth.[17,18] In our research, exposure from maternal smoking pregnancy was related to occurrence of COPD and also contributed to its severity. This is important information for COPD prevention. Third, the relationship between BMI and the incidence and severity of COPD is still under discussion, with no conclusion drawn for now.[19,20] However, in Northeast China, BMI was an important factor in both the screening and discriminate models of COPD. This importance may be related to the fact that people in the north are usually stronger than those in the south. Low BMI and particularly low fat-free mass are associated with worse outcomes,[21] which might explain different prognoses between the north and the south for COPD patients. Fourth, fuel exposure, heating style and resident type were considered to be important screening variables in our study, but removed in the COPD model established for the southern part of China. These indicated regional differences in COPD pathogenesis should be considered during COPD study.

**Table 12** The Confusion Matrix of Decision Tree (Training Set with Cross-Validation)

| Prediction / Truth | y3=1 | y3=2 | y3=3 |
|---|---|---|---|
| y3=1 | 170 | 3 | 1 |
| y3=2 | 11 | 73 | 15 |
| y3=3 | 4 | 25 | 58 |

**Table 13** The Confusion Matrix of Decision Tree (Test Set)

| Prediction / Truth | y3=1 | y3=2 | y3=3 |
|---|---|---|---|
| y3=1 | 42 | 2 | 0 |
| y3=2 | 3 | 20 | 2 |
| y3=3 | 0 | 8 | 13 |

The establishment of COPD-related models has been reported previously. Acute exacerbation is known to have a detrimental impact on COPD prognosis. Garcia-Aymerich et al[22] studied 340 patients with COPD and acute exacerbation at four tertiary hospitals in the Barcelona area of Spain. The study established a Cox proportional hazards model to obtain independent relative risks of readmission for patients with COPD. Furthermore, since the main characteristic of COPD is irreversible flow limitation (decreased $FEV_1$), ZafariZ[23] acquired data about 5594 patients and developed an individualized prediction model for $FEV_1$ in smokers with mild-to-moderate COPD. Su et al[24] implemented the prediction model for COPD among people more than 40 years old with respiratory symptoms and smoking history (≥20 pack-years). In contrast to these studies, which were mainly aimed at smokers, Chen et al[25] used the data of 4167 participants from the Framingham Offspring Cohort as an accurate tool to predict long-term lung function trajectories and the risk of airflow limitation in a general population using 20 common predictors. Further, Cui et al[26] established a discriminant-function model based on Bayes' Rule by stepwise discriminant analysis of the data from 243 patients with COPD and 112 non-COPD individuals in urban and rural communities and local primary care settings in Guangdong Province, China. However, these studies

established different models to assess $FEV_1$ or COPD by different methods. The optimal model is not known without statistical comparison. Melanie et al[27] conducted a detailed study of 30 articles in the 4481 COPD model records and found that only 4 studies were of good quality and included for review. During the analysis of these four studies, scientists discovered that the studies have significant differences in the included predictive indicators and the statistical methods selected. Guerra et al[28] analyzed 25 studies with 27 prediction models and found that only 3 models used high-quality statistical approaches. Therefore, our study established different models and did statistical comparisons to determine the optimal primary screening model and the best discrimination model to evaluate COPD in Northeast China. The verification process also showed the high effectiveness of these two models.

A limitation for this study was that we established COPD models for Northeast China instead of all of China. In view of the large regional differences between the north and the south such as the environment and weather, which is crucial in the development of COPD, we decided it was necessary to analyze risk factors and set models separately. It will be helpful to understand the different phenotypes of COPD.

In brief, COPD in Northeast China had special regional risk factors such as mother's smoking history during pregnancy, BMI, resident type, fuel exposure and current heating style. Among the primary screening and the discriminant models constructed with these high-risk factors, optimal models were a logistic regression model for primary screening and a decision tree model for discrimination. By using these models, doctors can easily primarily screen COPD and assess its severity, especially during COPD surveys in Northeast China.

## Ethics Statement

Procedures and experiment protocols were performed in accordance with the National Institute of Health Guide for Care and were approved by the Ethics Committee of China Medical University in accordance with the Declaration of Helsinki. All participants provided written informed consent.

## Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; gave final approval of the version to be published; and agree to be accountable for all aspects of the work.

## Funding

This work was supported by grants from the National Key Research and Development Program of China (No. 2018YFC1313600) _ and from the National Natural Science Foundation of China (No. 81670085).

## Disclosure

The authors declared that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China pulmonary health [CPH] study): a national cross-sectional study. *Lancet*. 2018;391(10313):1706–1717. doi:10.1016/S0140-6736(18)30841-9
2. Zhou Y, Zhong NS, Li X, et al. Tiotropium in early-stage chronic obstructive pulmonary disease. *N Engl J Med*. 2017;377(10):923–935. doi:10.1056/NEJMoa1700228
3. Dilektasli AG, Porszasz J, Casaburi R, et al. A novel spirometric measure identifies mild COPD unidentified by standard criteria. *Chest*. 2016;150(5):1080–1090. doi:10.1016/j.chest.2016.06.047
4. Mirsadraee M, Boskabady MH, Attaran D. Diagnosis of chronic obstructive pulmonary disease earlier than current Global initiative for obstructive lung disease guidelines using a feasible spirometry parameter (maximal-mid expiratory flow/forced vital capacity). *Chron Respir Dis*. 2013;10(4):191–196. doi:10.1177/1479972313507461
5. Caubet Fernandez M, Drouin S, Samoilenko M, et al. A Bayesian multivariate latent t-regression model for assessing the association between corticosteroid and cranial radiation exposures and cardiometabolic complications in survivors of childhood acute lymphoblastic leukemia: a PETALE study. *BMC Med Res Methodol*. 2019;19(1):100. doi:10.1186/s12874-019-0725-9
6. Vavougios GD, Doskas T, Konstantopoulos K. An electroglottographical analysis-based discriminant function model differentiating multiple sclerosis patients from healthy controls. *Neurol Sci*. 2018;39(5):847–850. doi:10.1007/s10072-018-3267-8
7. International Primary Care Airway Group. Ipag diagnosis management handbook–chronic airways disease. A guide for primarycare-physician[M / 0L]. (2005—1)[2009-3—15]. Available from: http://www.ipagguide.org.
8. Levy ML, Fletcher M, Price DB, et al. International Primary Care Respiratory Group (IPCRG) guidelines: diagnosis of respiratory diseases in primary care. *Prim Care Respir J*. 2006;15:20–34.
9. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. The St. George's respiratory questionnaire. *Am Rev Respir Dis*. 1992;145(6):1321–1327. doi:10.1164/ajrccm/145.6.1321
10. Fletcher CM. Standardised questionnaire on respiratory symptoms: a statement prepared and approved by the MRC Committee on the Aetiology of Chronic Bronchitis (MRC breathlessness score). *BMJ*. 1960;2:1662.
11. Jones PW, Harding G, Berry P, et al. Development and first validation of the COPD assessment test. *Eur Respir J*. 2009;34(3):648–654. doi:10.1183/09031936.00102509
12. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J*. 2005;26(2):319–338. doi:10.1183/09031936.05.00034805
13. for the PROBAST Group†; Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58. doi:10.7326/M18-1376.

14. McCloskey SC, Patel BD, Hinchliffe SJ, Reid ED, Wareham NJ, Lomas DA. Siblings of patients with severe chronic obstructive pulmonary disease have a significant risk of airflow obstruction. *Am J Respir Crit Care Med*. 2001;164(8):1419–1424. doi:10.1164/ajrccm.164.8.2105002
15. Ding Z, Wang K, Li J, Tan Q, Tan W, Guo G. Association between glutathione S-transferase gene M1 and T1 polymorphisms and chronic obstructive pulmonary disease risk: a meta-analysis. *Clin Genet*. 2018;95(1):53–62. doi:10.1111/cge.13373
16. Tager IB, Ngo L, Hanrahan JP. Maternal smoking during pregnancy. Effects on lung function during the first 18 months of life. *Am J Respir Crit Care Med*. 1995;152(3):977–983. doi:10.1164/ajrccm.152.3.7663813
17. Barker DJ, Godfrey KM, Fall C, Osmond C, Winter PD, Shaheen SO. Relation of birth weight and childhood respiratory infection to adult lung function and death from chronic obstructive airways disease. *BMJ*. 1991;303(6804):671–675. doi:10.1136/bmj.303.6804.671
18. Todisco T, de Benedictis FM, Iannacci L, et al. Mild prematurity and respiratory functions. *Eur J Pediatr*. 1993;152(1):55–58. doi:10.1007/BF02072517
19. Harikkhan RI, Fleg JL, Wise RA. Body mass index and the risk of COPD. *Chest*. 2002;121(2):370–376. doi:10.1378/chest.121.2.370
20. Liu Y, Pleasants RA, Croft JB, et al. Body mass index, respiratory conditions, asthma, and chronic obstructive pulmonary disease. *Respir Med*. 2015;109(7):851–859. doi:10.1016/j.rmed.2015.05.006
21. Guo Y, Zhang T, Wang Z, et al. Body mass index and mortality in chronic obstructive pulmonary disease: a dose-response meta-analysis. *Medicine (Baltimore)*. 2016;95(28):e4225. doi:10.1097/MD.0000000000004225
22. Garcia-Aymerich J, Farrero E, Félez MA, Izquierdo J, Marrades RM, Antó JM. Risk factors of readmission to hospital for a COPD exacerbation: a prospective study. *Thorax*. 2003;58(2):100–105. doi:10.1136/thorax.58.2.100
23. Zafari Z, Sin DD, Postma DS, et al. Individualized prediction of lung-function decline in chronic obstructive pulmonary disease. *Can Med Assoc J*. 2016;188(14):1004–1011. doi:10.1503/cmaj.151483
24. Su KC, Ko HK, Chou KT, et al. An accurate prediction model to identify undiagnosed at-risk patients with COPD: a cross-sectional case-finding study. *NPJ Prim Care Respir Med*. 2019;29(1):22. doi:10.1038/s41533-019-0135-9
25. Chen W, Sin DD, FitzGerald JM, Safari A, Adibi A, Sadatsafavi M. An individualized prediction model for long-term lung function trajectory and risk of COPD in the general population. *Chest*. 2019;157(3):547–557.
26. Cui J1, Zhou Y, Tian J, et al. A discriminant function model as an alternative method to spirometry for COPD screening in primary care settings in China. *J Thorac Dis*. 2012;4(6):594–600. doi:10.3978/j.issn.2072-1439.2012.11.06
27. Melanie M, Gayan B, Jennifer P, et al. Prediction models for the development of COPD: a systematic review. *Int J Chron Obstruct Pulmon Dis*. 2018;13:1927–1935. doi:10.2147/COPD.S155675
28. Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. *Eur Respir Rev*. 2017;26(143):pii:160061. doi:10.1183/16000617.0061-2016