A discriminant function model as an alternative method to spirometry for COPD screening in primary care settings in China

Jiangyu Cui¹*, Yumin Zhou¹*, Jia Tian¹, Xinwang Wang², Jingping Zheng¹, Nanshan Zhong¹, Pixin Ran¹

¹State Key Laboratory of Respiratory Diseases, Guangzhou Institute of Respiratory Diseases, The First Affiliated Hospital, Guangzhou Medical University, Guangzhou, China; ²Department of Statistics, Public Health School, Guangzhou Medical University, Guangzhou, China

ABSTRACT

Objective: COPD is often underdiagnosed in a primary care setting where the spirometry is unavailable. This study was aimed to develop a simple, economical and applicable model for COPD screening in those settings.

Methods: First we established a discriminant function model based on Bayes’ Rule by stepwise discriminant analysis, using the data from 243 COPD patients and 112 non-COPD subjects from our COPD survey in urban and rural communities and local primary care settings in Guangdong Province, China. We then used this model to discriminate COPD in additional 150 subjects (50 non-COPD and 100 COPD ones) who had been recruited by the same methods as used to have established the model. All participants completed pre- and post-bronchodilator spirometry and questionnaires. COPD was diagnosed according to the Global Initiative for Chronic Obstructive Lung Disease criteria. The sensitivity and specificity of the discriminant function model was assessed.

Results: The established discriminant function model included nine variables: age, gender, smoking index, body mass index, occupational exposure, living environment, wheezing, cough and dyspnoea. The sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, accuracy and error rate of the function model to discriminate COPD were 89.00%, 82.00%, 4.94, 0.13, 86.66% and 13.34%, respectively. The accuracy and Kappa value of the function model to predict COPD stages were 70% and 0.61 (95% CI, 0.50 to 0.71).

Conclusions: This discriminant function model may be used for COPD screening in primary care settings in China as an alternative option instead of spirometry.

KEY WORDS

COPD; Bayes’ Rule; spirometry


Introduction

Chronic obstructive pulmonary disease (COPD) has been predicted to become the fifth leading burden of disease in 2020 (1-3). Nevertheless, COPD is underdiagnosed (4) as most patients did not seek medical attention until they have serious respiratory symptoms. As reported in a recent Chinese population-based study (5), only 35.1% of the patients with “emphysema”, “asthma”, “bronchitis”, or “COPD” were identified by spirometry previously. Even in the U.S., 71.7% of the subjects with mild airflow limitation did not receive an appropriate diagnosis of obstructive lung disease (6).

Although spirometry is a “gold standard” for COPD detection, it is often underused in primary care settings, particularly in China (5) because it requires skills to operate and unfits for some patients. Our previous study (5) reported that only 6.5% of the patients with COPD had ever been tested by spirometry. Thus, it is of great value to develop a simple and economical method which can be used as an alternative option for spirometry to screen COPD and to predict the COPD stage in primary care settings. The aim of this study was to develop a mathematical model which can satisfy the above requirements.

Methods

Design

A total of 505 subjects (343 COPD patients and 162 non-COPD subjects), aged 40 years or over, were recruited from our previous...
population-based epidemiological study in the communities and local primary care settings in Shaoguan and Liwan, China. The protocol for the present population-based epidemiological survey was published elsewhere (7). The questionnaire and spirometry used for the present study was the same as used for COPD screening among outpatients at local primary care settings in 2008. According to the diagnostic criteria in Global Initiative for Chronic Obstructive Lung Disease (GOLD), COPD was diagnosed by a post-bronchodilator FEV1/FVC ratio < 0.7 measured after administration of albuterol. Non-COPD subjects were randomly selected from our previous population-based epidemiological data set using computer. All patients with COPD and Non-COPD subjects selected in this study for analysis except for those with a pre-existing or concomitant non-obstructive lung disease (e.g., pneumonophthisis, bronchiectasia, congestive heart failure, tuberculosis and lung cancer), those with acute respiratory symptoms, unstable hearing disease or other serious diseases, those with disability of walking due to other diseases, those without available data (from pre- and post-bronchodilator spirometric testing and a questionnaire) and those refusals.

The total participants were randomly split into two subsets: Training Set and Validation Set. The Training Set, consisting of 118 COPD patients at stages I-II, 125 COPD patients at stages III-IV and 112 non-COPD subjects, was used to establish the discriminant function model based on Bayes’ Rule. The Validation Set, including 150 subjects who had been randomly selected from the strata of non-COPD, COPD at stage I-II and COPD at stage III-IV, was used to evaluate the sensitivity, specificity and likelihood ratio of the established discriminant function model in COPD screening. The study protocol was approved by the Medical Ethic Committee at Guangzhou Institute of Respiratory Diseases and a written informed consent was given by all participants.

**Questionnaire**

The questionnaire used in this study was a revised form of the international BOLD study (8) and incorporated parts of the questionnaire was the same as used in our previous study in China (9). The questionnaire covered demographic data, respiratory symptoms/disease, comorbidities, health care use, activity limitation, nutritional status, potential risk factors for COPD, the Medical Research Council Dyspnoea Scale and health status (10). Occupational exposure was defined as exposure to any of noxious agents (dusts, chemicals, and gases) in any of the places where the subject had ever worked for at least 1 year. Smoking index was calculated by the pack number of smoking cigarettes each day multiplied years of smoking. The patients who had suffered from bovillae, pertussis or other respiratory infection in childhood were regarded as having a childhood infection history. Family history refers to a history of COPD in the family members like mother, father, brother or sister of the patient.

**Spirometry**

Spirometry was performed according to the American Thoracic Society (ATS) criteria (11) and ERS recommendations (12). Each spirometer was calibrated daily with a volume variation of less than 3% by a 3-L syringe. Spirometry operators had been well trained and accredited before the survey. The testing was repeated until three reproducible, acceptable results were obtained and the best FEV1, FVC, and FEV1/FVC ratios were recorded. Subjects with airflow limitations, which was defined as an FEV1/FVC ratio <70%, underwent post-bronchodilator testing at 15 to 20 minutes after inhaling a dose of 400 mcg of salbutamol (Ventolin; GlaxoSmithKline, Middlesex, UK) through a 500-mL spacer. An increase in FEV1 >12% and >200 mL from baseline was considered positive.

**Quality control**

Our quality control standard was reported previously (4). All interviewers had been well trained and accredited before the study began. A pre-investigation was conducted. Each completed questionnaire and spirometry report was verified. All questionnaire data were coded and entered into a standardised Excel database (Microsoft, Redmond, WA) by two independent investigators, with computer programs checking for out-of-range values and logic mistakes.

**Statistical analysis**

First, a discriminant function model based on Bayes’ Rule was established by stepwise discriminant analysis using the data from the Training Set. Our initial discriminant factors included the following variables reportedly associated with COPD: gender, index of smoking, body mass index, family history of COPD, educational history, child infection history, dyspnoea scale, occupational exposure, living environment, cooking fuels, wheezing, cough and cough with production (9,13-20). By stepwise discriminant analysis, the statistically significant variables were entered into the final discriminant function model and a retrospective discrimination was conducted among those individuals in the Training Set (21-38). Using the established model, the individuals in the Validation Set was then discriminated and the sensitivity, specificity and likelihood ratios of the model were assessed. We additionally evaluated the effect of the established model on discrimination of the COPD severity. All analyses were performed using the SAS version 9.1 software (SAS Institute, Cary, NC).
The COPD group included 215 males and 28 females, with ages ranging from 40 to 86 yrs. A total of 112 non-COPD individuals were recruited as a control group, including 51 females and 61 males, 40 to 79 yrs of age. A total of 84% of the study population were current smokers. Approximately 59% of the study population come from rural. Further characteristics of the enrolled subjects are given in Table 1.

Of the variables considered, nine were determined as significant discriminatory factors: age, gender, index of smoking, body mass index, occupational exposure, living environment, wheezing, cough and dyspnoea scale (Table 2).
Table 2. Variables and point values used for the model.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Variables</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living environment</td>
<td>× 1*</td>
<td>Urban =0, rural area =1</td>
</tr>
<tr>
<td>Gender</td>
<td>× 2*</td>
<td>Male =1, female =0</td>
</tr>
<tr>
<td>Age</td>
<td>× 3*</td>
<td>Years</td>
</tr>
<tr>
<td>Educational level</td>
<td>× 4</td>
<td>No =0, primary school =1, Junior school =2, high school =3, college =4, graduate =5</td>
</tr>
<tr>
<td>Dyspnoea</td>
<td>× 5*</td>
<td>Only get breathless with strenuous exercise =1; Get short of breath when hurrying on the level or up a slight hill =2; Walk slower than people of the same age on the level because of breathlessness or have to stop for breath when walking at one’s own pace on the level =3; Stop for breath after walking 100 yards or after a few minutes on the level = 4; Too breathless to leave the house =5</td>
</tr>
<tr>
<td>Cooking fuels</td>
<td>× 6</td>
<td>No =0, electricity =1, gas =2, coal =3, biomass fuel =4</td>
</tr>
<tr>
<td>Smoking index</td>
<td>× 7*</td>
<td>Number of packs/day*years</td>
</tr>
<tr>
<td>Occupational exposure</td>
<td>× 8*</td>
<td>Years</td>
</tr>
<tr>
<td>Family history</td>
<td>× 9</td>
<td>No =0, one person =1, two persons or more =2</td>
</tr>
<tr>
<td>Childhood infection history</td>
<td>× 10</td>
<td>No =0, one infection =1, two infections =2, three infections =3</td>
</tr>
<tr>
<td>Body mass index</td>
<td>× 11*</td>
<td>Weight(kg)/body height(meter)square kg/m²</td>
</tr>
<tr>
<td>Cough</td>
<td>× 12*</td>
<td>Cough for more than two years &amp; over three months/year no =0 yes = years</td>
</tr>
<tr>
<td>Sputum</td>
<td>× 13</td>
<td>Cough with production for more than two years &amp; over three months/year no =0 yes = years</td>
</tr>
<tr>
<td>Wheezing</td>
<td>× 14*</td>
<td>No =0, yes =1</td>
</tr>
</tbody>
</table>

*These statistically significant variables were incorporated into the final function model by stepwise discriminant analysis.

The equations are shown in the following box:

For control group:
\[
Y_0 = -59.01354 - 1.26683x_1 + 7.56887x_2 + 0.64493x_3 - 0.83960x_5 - 0.07024x_7 - 0.03545x_8 + 3.14363x_11 + 0.07232x_12 + 2.01073x_14
\]

For COPD patients of grade I/II:
\[
Y_1 = -58.27013 - 3.17714x_1 + 9.52916x_2 + 0.78278x_3 - 0.68573x_5 - 0.02098x_7 - 0.03143x_8 + 2.67730x_11 + 0.54014x_12 + 1.43502x_14
\]

For COPD patients of grade III/IV:
\[
Y_2 = -58.63480 - 3.29184x_1 + 9.73954x_2 + 0.81083x_3 - 0.27664x_5 - 0.01266x_7 - 0.00501x_8 + 2.48299x_11 + 0.59150x_12 + 2.79344x_14
\]

Therefore, given an individual’s values of \(x_1\), \(x_2\), \(x_3\), \(x_5\), \(x_7\), \(x_8\), \(x_{11}\), \(x_{12}\) and \(x_{14}\) from the questionnaire, we could calculate the values of \(Y_0\), \(Y_1\) and \(Y_2\) and then calculate their health status according to the highest values among \(Y_0\), \(Y_1\) and \(Y_2\) (based on Bayes’ rule). We retrospectively discriminated individuals in the Training Set using this model. As a result, the model had a sensitivity of 93.83%, a specificity of 89.29%, a positive likelihood ratio of 9.23, and a negative likelihood ratio of 0.07, an accuracy of 92.4% and an error rate of 7.6% in retrospective discrimination. Next, 150 individuals were discriminated by the functions of sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, accuracy and error rates of 89.00%, 82.00%, 4.94, 0.13, 86.66% and 13.34%, respectively (Table 3). The discriminant function model resulted in an accuracy of 74.09% and a Kappa value of 0.70 (95% CI, 0.65 to 0.76) for retrospective prediction of COPD stage, as well as an accuracy of 70% and Kappa value of 0.61 (95% CI, 0.50 to 0.71) (Table 4).

Discussion

Our study tentatively developed a discriminant function model consisting of nine variables which can be applied to screen COPD as an alternative option in areas where spirometry is unavailable. To our knowledge, no previous study used a discriminant function model to screen for COPD, though the same way has been used in the diagnosis of other diseases (39). The present studies demonstrated clearly that spirometry underused in primary care settings not only in China but also in other developing countries. Previous studies developed questionnaires as a diagnostic scoring system of COPD and used them to identify persons who are likely to have COPD among...
We devised a discriminant function model to diagnose COPD according to the patient’s answers to some simple questions in the questionnaire. The model was demonstrated in the present study to have such high sensitivity (>89.00%) and specificity (>82.00%) that it can be used in primary care settings in China, especially in the rural areas, where spirometry is unavailable. A doctor could easily diagnose COPD using our patient questionnaire and software-based calculations by the model.

Compared with spirometry, the short screening questionnaire of nine variables is much simpler, easier and economical. It seems to be a more sensitive method to screen COPD than the scoring system of the COPD diagnostic questionnaires which were reported to have sensitivities of 54% to 82% and specificities of 58% to 88% (42). In addition, our discriminant function model can also be used to predict the stage of COPD, with an accuracy of about 70%.

It is well known that the discriminatory effects of a mathematical model depend on the variables selected. We selected initially 14 variables probably associated with COPD (see Table 2) according to the published literature (9-20) to detect the risk factors of COPD by stepwise screening (21-36). At last, nine variables were identified as discriminatory factors and were devised to a discriminant function model to screen COPD by some simple questions. In our discriminant function model, both smoking and BMI are the most significant discriminatory factors, which is consistent with literature. It is well-known that smoking is considered the most important risk factor in the development

### Table 3. Diagnosis of COPD by spirometry vs. by outcome predicted by the model.

<table>
<thead>
<tr>
<th>Discriminant by model</th>
<th>Diagnosis by spirometry*</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COPD</td>
<td>Non-COPD</td>
</tr>
<tr>
<td>Training Set†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted COPD</td>
<td>228</td>
<td>12</td>
</tr>
<tr>
<td>Predicted non-COPD</td>
<td>15</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>243</td>
<td>112</td>
</tr>
<tr>
<td>Validation Set‡</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted COPD</td>
<td>89</td>
<td>9</td>
</tr>
<tr>
<td>Predicted non-COPD</td>
<td>11</td>
<td>41</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>50</td>
</tr>
</tbody>
</table>

*Diagnosis by spirometry was performed using the criteria of FEV1/FVC <70%; † In Training Set that were used to establish the model by stepwise discriminant analysis: sensitivity =93.83%, specificity =89.29%, accuracy =92.4%, PPV =95%, NPV =86.96%, +LR =9.23, -LR =0.07; ‡ In Validation Set that were used to check the model: sensitivity =89.00%, specificity =82.00%, accuracy =86.66%, PPV =90.82%, NPV =78.85%, +LR =4.94, -LR =0.13.

### Table 4. Diagnosis of COPD and stage by spirometry* vs. by outcome predicted by the model.

<table>
<thead>
<tr>
<th>Discriminant by model</th>
<th>Diagnosis by spirometry*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-COPD</td>
</tr>
<tr>
<td>Training Set†</td>
<td></td>
</tr>
<tr>
<td>Predicted non-COPD</td>
<td>100</td>
</tr>
<tr>
<td>Predicted grade I/II</td>
<td>11</td>
</tr>
<tr>
<td>Predicted grade III/IV</td>
<td>1</td>
</tr>
<tr>
<td>Validation Set‡</td>
<td></td>
</tr>
<tr>
<td>Predicted non-COPD</td>
<td>41</td>
</tr>
<tr>
<td>Predicted grade I/II</td>
<td>4</td>
</tr>
<tr>
<td>Predicted grade III/IV</td>
<td>5</td>
</tr>
</tbody>
</table>

*Diagnosis by spirometry was performed using the criteria of FEV1/FVC <70%; † In Training Set that were used to establish the model by stepwise discriminant analysis: accuracy =74.09%, Kappa =0.70 (95% CI, 0.65 to 0.76); ‡ In Validation Set that were used to check the model: accuracy =70%, Kappa =0.61 (95% CI, 0.50 to 0.71).
of COPD. In China, about two-thirds (61.4%) of the patients with COPD, 81.8% among the male patients and 24.0% among the female patients, were smokers; 13.2% of the smokers had COPD and the risk for COPD increased with the number of cigarettes consumed. In Korea, 88% of the male patients with COPD were smokers, and 36% of the adult smokers (45 years of age or older) who had smoked at least 20 cigarettes/day were diagnosed with COPD. Since BMI is another most important risk factor for COPD, those with a BMI of less than 18.5 kg/m² may have a COPD prevalence of as high as 21.0% and there is a negative correlation between BMI and COPD prevalence. However, some risk factors, such as use of cooking fuels, sputum production, childhood infection, educational level, and familial history, were removed from the model of ours, mainly because the regions involved in the present investigation are highly correlated with the usage of cooking fuels and educational level is strongly correlated with both region and age.

Some limitations should be noted in this study. First, the sample size of 355 participants in this study may be insufficient for characterization of an entire population. Secondly, not enough women were recruited for the present study, possibly because the morbidity of COPD in women is lower than in men. Thirdly, we identified patients with COPD according to the GOLD criteria which might have led to overdiagnosis of COPD in older people. In addition, as the discriminant function model was only used as a screening tool rather than as a diagnostic criterion, the COPD patients identified by the model should have been confirmed by the spirometry. At last, although the discriminant function model was developed from the data of the subjects in Guangdong, the nine variables associated with COPD were generalised from the data of the subjects beyond Guangdong. The discrepancy may have had an unknown influence on the efficacy of the model.

In conclusion, the discriminant function model reported here is a first attempt of its kind to develop an alternative method for the COPD screening in Chinese settings. We believe that it may help diagnosis early enough a great number of COPD patients who may not be diagnosed otherwise and the early diagnosis can allow them to have a timely medical treatment.

**Acknowledgements**

Funding: This study was supported by the National Key Technology R&D Program of the 12th National Five-year Development Plan (2012BAI05B01). The funding providers, however, had no influence on the study design, collection, analysis, interpretation of the data, writing of the report and the decision to submit the report for publication.

Contributors: JC collected the data, monitored data collection, planned the statistical analysis, analyzed the data, and drafted and revised the manuscript. YZ and JT collected the data and revised the manuscript. PR and NZ initiated and designed the project, monitored data collection, drafted and revised the manuscript. PR and JC are guarantors.

Competing interests: all authors have completed the Unified Competing Interest Form at www.icmje.org/coi_disclosure. pdf (available on request from the corresponding author), and declared herein no support from any institution for the submitted work, no financial relationship with any institution that might have an interest in the submitted work in the previous 3 years and no other relationship or activity that could appear to have influenced the submitted work.

Ethical approval: the study protocol was approved by Medical Ethic Committee at Guangzhou Institute of Respiratory Diseases and a written informed consent was given by all participants.

**References**

Cui et al. Spirometry for COPD screening in primary care settings


