





# Evaluation of Risk Factors for Chronic Obstructive Pulmonary Disease in the Middle-Aged and Elderly Rural Population of Northeast China Using Logistic Regression and Principal Component Analysis

Rui Wang <sup>1-4</sup>, Wei Zhang <sup>1-3</sup>, Yuanyuan Li <sup>1-3</sup>, Yuting Jiang <sup>1-3</sup>, Hongqi Feng <sup>1-3</sup>, Yang Du <sup>1-3</sup>, Zhe Jiao <sup>1-3</sup>, Li Lan <sup>4</sup>, Xiaona Liu <sup>1-3</sup>, Bingyun Li <sup>1-3</sup>, Chang Liu <sup>1-3</sup>, Xingbo Gu <sup>5</sup>, Fang Chu <sup>1-3</sup>, Yuncheng Shen <sup>1-3</sup>, Chenpeng Zhu <sup>1-3</sup>, Xinhua Shao <sup>1-3</sup>, Simeng Tong <sup>1-3</sup>, Dianjun Sun <sup>1-3</sup>

<sup>1</sup>Center for Endemic Disease Control, Chinese Center for Disease Control and Prevention, Harbin Medical University, Harbin, 150081, People's Republic of China; <sup>2</sup>National Health Commission & Education Bureau of Heilongjiang Province, Key Laboratory of Etiology and Epidemiology, Harbin Medical University (23618504), Harbin, 150081, People's Republic of China; <sup>3</sup>Heilongjiang Provincial Key Laboratory of Trace Elements and Human Health, Harbin Medical University, Harbin, 150081, People's Republic of China; <sup>4</sup>Harbin Center for Disease Control and Prevention, Harbin, 150056, People's Republic of China; <sup>5</sup>Department of Biostatistics, School of Public Health, Hainan Medical University, Haikou, 571199, People's Republic of China

Correspondence: Dianjun Sun, Center for Endemic Disease Control, Chinese Center for Disease Control and Prevention, Harbin Medical University, Baojian Road 157, Harbin, Heilongjiang Province, 150081, People's Republic of China, Tel +86-13313685318, Fax +86 45186612695, Email hrbmusdj@163.com

**Purpose:** To investigate the environmental, immune, and inflammatory factors associated with chronic obstructive pulmonary disease (COPD) in middle-aged and older Chinese individuals.

**Patients and Methods:** A community-based case-control study was conducted among 471 patients with COPD and 485 controls. The information on COPD of the participants was collected through face-to-face interviews, and serum samples were measured at the laboratory. The main risk factors for COPD were analyzed using principal component analysis (PCA) and logistic regression.

**Results:** Nine hundred and fifty-six respondents were included in the analysis. The results of the PCA-logistic regression analysis showed significant differences in the environmental factors, medical history, and serum C-reactive protein (CRP) levels between patients and controls. COPD was markedly more usual in those with smoking index >200 (OR, 1.42; 95% CI, 1.28–1.57); exposure to outdoor straw burning (OR, 1.64; 95% CI, 1.47–1.83); use of coal, wood, and straw indoors (OR, 2.31; 95% CI, 1.92–2.78); history of respiratory disease and coronary heart disease (OR, 3.58; 95% CI, 3.12–4.10), congestive heart failure (OR, 1.23; 95% CI, 1.09–1.38), and cerebrovascular disease (OR, 1.15; 95% CI, 1.02–1.31); and higher serum level of CRP (OR, 1.20; 95% CI, 1.11–1.30). Compared to the logistic regression analysis, PCA logistic regression analysis identified more important risk factors for COPD.

**Conclusion:** PCA-logistic regression analysis was first utilized to explore the influencing factors among rural residents in Northeast China Environmental aged 40 years and above, it was found that environmental factors, medical history, and serum CRP levels mainly affected the prevalence of COPD.

**Keywords:** chronic obstructive pulmonary disease, collinearity, principal component, logistic regression

## Introduction

Chronic obstructive pulmonary disease (COPD) is defined as a treatable and preventable disease. Its pulmonary component is characterized by an irreversible and progressive airflow limitation due to an abnormal inflammatory response of the lung to noxious gases or particles.<sup>1</sup> It is often accompanied by significant extra pulmonary effects, leading to the aggravation of the disease. In the early onset of the disease, people with COPD may develop shortness of breath during exercise. As the disease advances, it may be difficult to inhale or even exhale. Persons with COPD may have emphysema, bronchiolitis, or both. The magnitude of each component varies from person to person.<sup>2</sup>

Tobacco smoking is the leading cause of COPD, though not the only one. Some pathobiological processes interact in combination with lung growth, environmental stimuli, and genetic factors. Another characteristic of COPD is an infection-induced deterioration, with up to 78% of patients with a severe disease being caused by viral infection, bacterial infection, or both. COPD is also accompanied by several comorbidities, including lung cancer, cardiac disease, and diabetes.<sup>3</sup> The main feature of COPD is chronic progressive inflammation of the lung tissue. However, systemic immune and inflammatory responses are also important features of COPD.

COPD is considered as one of the major reasons of mortality and morbidity worldwide, causing heavy healthcare, economic, and social burdens.<sup>4</sup> The prevalence of COPD differs between countries, reflecting the diverse age composition of the population, genetic predispositions, healthcare resources, and various exposure to risk factors.<sup>5</sup> The prevalence of COPD is almost 4% of the global population.<sup>6</sup> Prevalence rises to 10% among people aged 40 and older.<sup>7</sup> Air pollution caused by over 40% of the global population using biomass as indoor fuel accounted for 7.7% of the world's mortality in 2012.<sup>8</sup> In low- and middle-income countries, exposure to biomass smoke accounts for approximately 25% of premature deaths from COPD.<sup>9</sup> In 2016, COPD was the fifth leading risk factor for death in China.<sup>10</sup> The general prevalence of COPD in China was 8.2%. The risk of COPD and its complications increases with increasing age, especially in those over 40 years of age. Fang et al in 2014–2015 noted that in China 13.6% of adults aged 40 years and above had COPD and its prevalence in rural areas was higher than that in urban areas. The prevalence of COPD varies substantially between geographic areas in China and was found to be the second highest in northeast China.<sup>11</sup> Countries with lower income levels and low-income people usually have a greater risk of COPD. Most previous studies evaluating influencing factors have aimed at the entire population or the urban population. The large gap in economic, societal practices and healthcare in urban and rural areas accounts for the diverse risk factors for COPD. At present, there is still a lack of research on the factors influencing COPD in people over 40 years of age in the rural Northeast areas of China.

Logical regression is commonly used for risk factor analysis.<sup>12</sup> However, if there is collinearity between the variables, the decreased stability of the regression equation leads to a greatly lower quality of the prediction results.<sup>13</sup> Principal component analysis (PCA) is a method of linear dimensionality reduction and feature extraction by streamlining sets of variables from large to small, while still including a chunk of the primitive information.<sup>14</sup> The PCA approach is to seek a linear combination of the principal component (PC) to maximize the variance, then remove this variance and seek a linear combination of the second principal component, iterating until the principal component is sufficient to represent most of the primitive information. Moreover, the eigenvalues and the eigenvectors of the covariance matrix must be calculated, where the eigenvalues are used to do the descending order arrangement, and the eigenvectors project the actual data on the basis of the ranking. PCA and logistic regression often play an important role in identifying associated risk factors and people at high risk for COPD.<sup>15–17</sup>

Populations in rural areas of Northeast China have environmental characteristics such as smoking, exposure to outdoor straw burning and use of fuel indoors, and residents over the age of 40 have high levels of comorbidity, systemic immunity and inflammatory responses. Our study was conducted to consider the main etiological factors and the importance of each factor in people aged 40 years and above in the rural areas of Northeast China and first use PCA-logistic regression analysis to explore the impact of COPD.

## Materials and Methods

### Selection of Cases and Controls

A case–control study was conducted on the community population. Based on the baseline data of a 10,000-person cohort study, cluster sampling was performed from five villages in Mingshui County, Suihua City, Heilongjiang Province, China, and a total of 566 COPD cases and 566 controls were randomly sampled. A case–control study was carried out, and epidemiological investigations and general physical conditions were carried out on the villagers. A total of 1132 people were investigated, excluding 176 people who had incomplete epidemiological data for telephone follow-up and could not be supplemented. In the end, this study consisted of 956 people aged 40–76 years. Lung function measurements were performed by trained personnel at the rural community health centers. COPD was confirmed in patients with a ratio of forced expiratory volume in one second to forced vital capacity of <70% ( $FEV_1/FVC < 70\%$ ) to represent the definition

of the epidemiologic case. Individuals with a FEV<sub>1</sub>/FVC ratio of less than 0.7 were selected as cases, and the remaining were eligible as controls. The study was in accordance with the Declaration of Helsinki and was conducted after obtaining approval from the Ethics Committee of Harbin Medical University. Each patient signed a written informed consent.

## Questionnaires and Serological Evaluation

The first face-to-face questionnaire-based interviews, fasting blood measurements, and anthropometric measurements were conducted at a rural community health center for both cases and controls. The second questionnaire-based survey was conducted via telephone. The investigators were trained to conduct face-to-face and telephonic interviews using the designed questionnaires. The questionnaire covered environmental information and basic demographics such as income, age, marital status, education level, clinical signs and symptoms, combinations, lifestyle behaviors, history of diseases, and psychosocial information. Before breakfast, qualified nurses gathered fasting blood samples through venipuncture. Serum samples were stored at  $-80^{\circ}\text{C}$ . Serum C-reactive protein (CRP), Immunoglobulin (Ig) G (IgG), IgM, IgA, the third complement component (C3), and the fourth complement component (C4) levels were measured using a Hitachi 3100 automatic biochemical analyzer at the central laboratory of the Endemic Disease Control Center, China CDC.

## Statistical Analyses and Sample-Size Evaluation

All data were gathered using dedicated software and analyzed applying SAS statistical analysis system in version 9.1.4. First, univariate analysis containing Wilcoxon rank sum test or *t*-test for continuous data and chi-square test for classification variables were performed to identify factors significantly associated with COPD. Variables with statistically significant results from the univariate analyses were selected for correlation analysis and collinear diagnosis. Finally, these variables were entered into the multivariate logistic regression and PCA-logistic models to obtain the odds ratio (OR) and 95% confidence interval (CI) for each risk factor.

The purpose of the sample-size evaluation was to find correlations (at least  $\text{OR} = 2$ ) between COPD and other variables, with a power of 90% and significance at 5%. Based on earlier researches, we set 20% of patients to reach the main findings of environmental exposure. Meanwhile, two-sided alpha levels were set to 0.05. We determined that 256 patients would provide a power of more than 90% to examine a diversity between the two groups. The sample capacity selected in this research was bigger than expected.

## Results

We compared the characteristics of 471 cases and 485 controls with respect to the main demographic characteristics, disease history, environmental exposure, and serum indices (Table 1). Statistical discrepancy in age, level of income, education, Body mass index (BMI), smoking status, indoor fuel use (no fuel/natural gas/electricity vs coal/wood/straw), outdoor straw burning, history of respiratory disease (RD), history of coronary heart disease and congestive heart failure (CHD/CHF), history of cerebrovascular disease (CVD), IgG, IgA, and CRP were found between the two groups ( $P < 0.05$ ).

To further explore the presence of a nonlinear association of IgM, C3, and C4 with COPD, the three variables were logarithmically transformed, and logistic nonlinear regression was used to analyze the role of quadratic effects. The results are presented in Table 2. The discrepancy in the C3<sup>2</sup> index between the two groups was statistically marked ( $P < 0.05$ ), while the difference in the IgM and C4 indices was not statistically noticeable ( $P > 0.05$ ).

## Correlation and Collinearity Analyses

The results of the correlation analysis of the main demographics, disease history, environmental exposure, and immune and inflammatory levels are presented in Table S1 in the Supplementary Appendix. Immune- and inflammation-related indicators, including IgG, IgA, CRP, C3, and C3<sup>2</sup> were markedly interrelated each other. Each other ( $P < 0.05$ ), and there were significant correlations between the main demographics, disease history, environmental exposures, and levels of immunity and inflammation ( $P < 0.05$ ). Owing to the obvious correlation, possible collinearity among the indicators was considered. The results of the collinearity diagnosis and variance inflation factors of the main demographics, disease

**Table 1** Main Demographics, Disease History, Environmental Exposure, and Serum Indices Between Patients with Chronic Obstructive Pulmonary Disease and Controls

Characteristics n (%) / (Median, Range)	FEV1/FVC<70% (471)	FEV1/FVC≥70% (485)	X <sup>2</sup> /t/Z-value	P-value
Sex	471	485	0.4902	>0.05
Male	186 (39.49)	181 (37.32)		
Female	285 (60.51)	304 (62.68)		
Age	59.41±8.14	57.47±8.22	-3.66	<0.05
BMI (kg/m <sup>2</sup> )	468	482	13.1033	<0.05
<18.5	28 (5.98)	9 (1.87)	8.0837	<0.05
[18.5, 24)	189 (40.38)	180 (37.34)	RG	
[24, 28)	176 (37.61)	202 (41.91)	1.6218	>0.05
>28	75 (16.03)	91 (18.88)	1.6704	>0.05
Level of income	471	485	11.7690	<0.05
<12,000	367 (77.92)	350 (72.16)	RG	
(12,000, 20,000)	65 (13.8)	60 (12.37)	0.0283	>0.05
>20,000	39 (8.28)	75 (15.46)	11.3430	<0.05
Level of education	471	485	11.3755	<0.05
Below primary school	140 (29.72)	108 (22.27)	RG	
Primary school/Junior high school	305 (64.76)	329 (67.84)	4.9654	<0.05
High school and above	26 (5.52)	48 (9.9)	10.3690	<0.05
Smoking status	441	456	13.5708	<0.05
Smoking index*≤200	240 (54.42)	303 (66.45)		
Smoking index>200	201 (45.58)	153 (33.55)		
Second-hand smoking (never-smokers)	150	310	0.5807	>0.05
Yes	70 (46.67)	133 (52.90)		
No	80 (53.33)	177 (57.10)		
Indoor fuel use	450	471	18.5861	<0.05
No fuel/Natural gas/Electricity	18 (4.00)	55 (11.68)		
Coal/Wood/Straw	432 (96.00)	416 (88.32)		
Outdoor straw burning	449	470	10.7516	<0.05
Yes	336 (74.83)	305 (64.89)		
No	113 (25.17)	165 (35.11)		
History of RD	471	485	51.9267	<0.05
Yes	112 (23.78)	34 (7.01)		
No	359 (76.22)	451 (92.99)		
History of CHD/CHF	471	485	4.2892	<0.05
Yes	127 (29.96)	103 (21.24)		
No	344 (73.04)	382 (78.76)		
History of CVD	471	485	4.6860	<0.05
Yes	108 (22.93)	84 (17.32)		
No	363 (77.07)	401 (82.68)		
IgG	16.01±4.29	16.69±4.37	2.37	<0.05
IgA	2.22±1.03	2.31±0.93	2.0000	<0.05
IgM	1.17±0.58	1.12±0.57	0.1930	>0.05
C3	1.22±0.24	1.24±0.56	-1.1509	>0.05
C4	0.30±0.09	0.30±0.11	-0.51	>0.05
CRP	3.69±6.52	3.02±4.75	2.5667	<0.05

**Notes:** Continuous variables are expressed as mean and standard deviation, while categorical variables are presented as the number of patients in each group and percentage in the study group with respect to the total population. \*Smoking index = the number of cigarettes smoked per day × the number of years of smoking.

**Abbreviations:** BMI, body mass index; RG, reference group; RD, respiratory disease; CHD/CHF, coronary heart disease and congestive heart failure; CVD, cerebrovascular disease; IgG, Immunoglobulin G; IgM, Immunoglobulin M; IgA, Immunoglobulin A; C3, the third complement component; C4, the fourth complement component; CRP, C-reactive protein.

**Table 2** Univariate Logistic Nonlinear Regression Model to Analyze the Effects of IgM, C3, C4 on Chronic Obstructive Pulmonary Disease

Characteristics	$\beta$	SE ( $\beta$ )	Wald Chi-Square Value	P-value
IgM	0.2969	0.3051	0.9474	0.3304
IgM2	-0.0477	0.0906	0.2777	0.5982
C3	3.5619	1.8539	3.6913	0.0547
C32	-1.5399	0.7252	4.5093	0.0337
C4	0.4791	0.4112	1.3577	0.2439
C42	-1.4583	0.9776	2.2255	0.1357

**Abbreviations:** IgM, Immunoglobulin M; C3, the third complement component; C4, the fourth complement component.

history, environmental exposure, and immune and inflammatory levels are shown in [Tables S2](#) and [S3](#) in the Supplementary Appendix. The variation inflation factors of the potential predictors ranged from 1.057 to 47.565, which was higher than the threshold of 10, indicating multicollinearity between the potential predictors.

## Logistic Regression Analysis

Applying multivariate gradual logistic regression to find variables associated with COPD. Univariate analysis showed that 15 variables were statistically marked. As shown in [Table 3](#), a smoking index >200, history of respiratory disease, indoor fuel use, outdoor straw burning, and higher serum levels of CRP and C3 were risk factors for COPD, while C3<sup>2</sup> was a protective factor for COPD; all these variables were statistically marked ( $P < 0.05$ ). The 95% confidence interval of C3 was too wide; thus, the results are not ideal and there may have been errors.

## Principal Component Analysis

To include as many risk factors for COPD as possible, all elements with  $P < 0.05$  were brought in PCA, for a total of 15 variables. As shown in [Table 4](#), we found that in the dataset the top 11 principal components (PCs) generated by the model interpreted 88% of the variance, and the top 14 PCs interpreted 99.93% of the variance. The top 11 and 14 PCs were used as independent variables, and presence of COPD (0 = no; 1 = yes) was used as the dependent variable in the multivariate gradual logistic regression model. The results were consistent and are listed in [Table 5](#), and the results of the feature vectors are listed in [Table S4](#) in the Supplementary Appendix. Among the various PCs, PC2, PC5, PC6, PC7 and PC8 were statistically significant ( $P < 0.05$ ). The loadings represent the level of significance of the corresponding compounds. The first three levels of significance of PC2 in order were age > indoor fuel use > CHD/CHF; the first three levels of significance of PC5 in order were smoking status > straw burning outdoors > CRP; the first three levels of significance of PC6 in order were education > RD > income; the first three levels of significance of PC7 in order were age > CVD > smoking status; and the first three levels of significance of PC8 in order were BMI > IgA > smoking status. The results of logistics regression

**Table 3** Multivariate Gradual Logistic Regression Analysis for Risk Factors of Chronic Obstructive Pulmonary Disease

Characteristics	$\beta$	SE ( $\beta$ )	OR (95% CI)	P-value
RD	1.5301	0.2331	4.619 (2.925–7.293)	0.0003
Smoking status	0.3976	0.1541	1.488 (1.100–2.013)	<0.0001
Indoor fuel use	0.8545	0.3127	2.350 (1.273–4.338)	0.0099
Straw burning outdoor	0.4244	0.1668	1.529 (1.103–2.120)	0.0063
CRP	0.3783	0.1341	1.460 (1.122–1.899)	0.0109
C3	5.7883	2.1593	326.442 (4.741–>999.999)	0.0048
C3 <sup>2</sup>	-2.5193	0.8505	0.081 (0.015–0.426)	0.0073

**Abbreviations:** RD, respiratory disease; C3, the third complement component; CRP, C-reactive protein.

**Table 4** Results of the Principal Component Analysis

Principal Component	Eigenvalue	Difference	Proportion	Cumulative
PRIN 1	2.65861343	0.71186102	0.1772	0.1772
PRIN 2	1.94675241	0.59745936	0.1298	0.3070
PRIN 3	1.34929305	0.17051207	0.0900	0.3970
PRIN 4	1.17878098	0.08205982	0.0786	0.4756
PRIN 5	1.09672117	0.09870783	0.0731	0.5487
PRIN 6	0.99801334	0.07258332	0.0665	0.6152
PRIN 7	0.92543002	0.07846652	0.0617	0.6769
PRIN 8	0.84696350	0.05007724	0.0565	0.7334
PRIN 9	0.79688626	0.05579577	0.0531	0.7865
PRIN 10	0.74109050	0.02792364	0.0494	0.8359
PRIN 11	0.71316686	0.05375892	0.0475	0.8834
PRIN 12	0.65940794	0.08171214	0.0440	0.9274
PRIN 13	0.57769580	0.07713305	0.0385	0.9659
PRIN 14	0.50056275	0.48994074	0.0334	0.9993
PRIN 15	0.01062200		0.0007	1.0000

**Table 5** Multivariate Gradual Logistic Regression Analysis for Principal Components of Chronic Obstructive Pulmonary Disease

Principal Component	$\beta$	SE ( $\beta$ )	Wald Chi-Square Value	P-value
Intercept	-0.0423	0.0731	0.3338	0.5634
PRIN 2	0.4343	0.0579	56.3579	<0.0001
PRIN 5	0.1468	0.0705	4.3315	0.0374
PRIN 6	0.2697	0.0749	12.9794	0.0003
PRIN 7	-0.243	0.0783	9.6322	0.0019
PRIN 8	0.2445	0.0807	9.1818	0.0024

analysis for the conversion of principal components into primary variables are shown in Table 6. Age, education, smoking status, indoor fuel use, outdoor straw burning, RD, CHD/CHF, CVD, and CRP were found to be marked ( $P < 0.05$ ); the remaining variables were not marked ( $P > 0.05$ ).

**Table 6** Results of Logistics Regression Analysis for Conversion of Principal Components into Primary Variables

The Primary Variable	$\beta$	SE ( $\beta$ )	U-value	P-value	OR (95% CI)
Age	0.0116	0.0031	3.7821	<0.05	1.0117(1.0056–1.0178)
BMI	-0.0195	0.0217	0.8974	>0.05	0.9807(0.9399–1.0233)
Level of education	-0.2532	0.0461	5.4954	<0.05	0.7763(0.7093–0.8497)
Level of income	0.0067	0.0367	0.1815	>0.05	1.0067(0.9368–1.0818)
Smoking status	0.3490	0.0517	6.7555	<0.05	1.4177(1.2811–1.5687)
Indoor fuel use	0.8367	0.0947	8.8379	<0.05	2.3088(1.9178–2.7796)
Outdoor straw burning	0.4935	0.0558	8.8407	<0.05	1.6381(1.4683–1.8275)
History of RD	1.2750	0.0693	18.4027	<0.05	3.5787(3.1243–4.0992)
History of CHD/CHF	0.2050	0.0588	3.4873	<0.05	1.2275(1.0939–1.3773)
History of CVD	0.1428	0.0637	2.2432	<0.05	1.1535(1.0182–1.3068)
IgG	-0.0085	0.0057	1.4759	>0.05	0.9916(0.9805–1.0028)
IgA	0.0789	0.0592	1.3328	>0.05	1.0821(0.9635–1.2153)
C3	0.0133	0.0995	0.1342	>0.05	1.0134(0.8339–1.2316)
C3 <sup>2</sup>	0.0059	0.0386	0.1521	>0.05	1.0059(0.9326–1.0849)
CRP	0.1815	0.0387	4.6893	<0.05	1.1990(1.1114–1.2935)

**Abbreviations:** BMI, body mass index; RD, respiratory disease; CHD/CHF, coronary heart disease and congestive heart failure; CVD, cerebrovascular disease; IgG, Immunoglobulin G; IgA, Immunoglobulin A; C3, the third complement component; CRP, C-reactive protein.

## Discussion

There is large variation in the reported prevalence of COPD among different geographical areas in China, partly due to the different levels of exposure to risk factors and disparities in socioeconomic development among different areas.<sup>18</sup> The overall trend in previous studies is that COPD is generally seen in men.<sup>11,19–22</sup> However, in this study, sex was found to be evenly balanced between patients with COPD and controls. The reason may be the higher rate of smoking in rural women compared with those in urban areas and in the entire population, and the high prevalence of occupational dust and indoor fuel pollution in rural districts. The enhanced risk of COPD with age is consistent with previous studies, possibly because of decreased lung function with age and the accumulated effect of environmental pollutants on the lungs.<sup>20,22</sup> Senile emphysema is presented with the damage of alveolar septum, the decline of elastic function, and the enlargement of pulmonary cavity, whose severity can be evaluated by expiratory flow limitation.<sup>23,24</sup> Lung aging can also lead to an abnormal immune response after infection or injury, increasing susceptibility to infection, and subsequently increasing lung injury.<sup>25</sup> The increased risk of COPD in people with low education may be due to poorer living conditions, lesser health education, and greater exposure to environmental pollutants.<sup>20,22–26</sup>

Environmental pollutants are generally considered the major pathogenic elements of COPD, identical with the findings of our study. Previous studies on the association between smoking and COPD have generally reported that current smoking, former smoking, and longer years of smoking increase the risk of developing COPD.<sup>20,21,27–31</sup> We divided the smoking index, a comprehensive index used to assess the severity of smoking, into two categories, and found a significantly greater risk of COPD with a smoking index >200, which has not been reported in previous domestic studies. Compared to the residents in urban areas, rural residents in Northeast China Environmental aged 40 years and above are less educated, and they tend to pay less attention to the harm of smoking, which leads to their long-term smoking behavior and makes it more difficult for them to quit smoking. Therefore, smoking index >200 being used as a risk predictor of COPD will easily and intuitively guide residents to reduce their smoking volume, which is more acceptable than direct smoking cessation, so as to achieve the ultimate goal of smoking cessation to promote health. We found that indoor fuel use, mainly firewood, coal, and straw, increased the risk of COPD, which is generally identical with findings of earlier researches.<sup>20,22,32</sup> Due to its geographical location, the rural areas of northeast China witness cold climates for large parts of the year. Due to rural living conditions, doors and windows are often closed in the cold weather, resulting in the retention of harmful gases indoors that are released after fuel combustion, thus increasing the risk of COPD. We also found that exposure to outdoor straw burning increased the risk of COPD. Although there are some reports focusing on exposure to biomass smoke exposure for COPD, there is still a lack of specific research on outdoor straw burning influencing COPD. The phenomenon of outdoor straw burning has improved due to appropriate government policies in recent years, but outdoor straw burning has become a pathogenic elements for COPD due to the accumulation of exposure for several years. COPD is a complex disease whose pathogenesis is interacted by multiple factors, including exposure to harmful particles and gases, genetic predisposing elements, and respiratory immune responses. Tobacco smoking is the leading cause of COPD in earlier researches. Hazardous gas inhalation can give rise to the injury of vascular endothelium and airway epithelium through premature aging, lung inflammation, impaired immune function, oxidative stress, and abnormal metabolic enzymes.<sup>33,34</sup>

This study found that the history of RD, CHD/CHF, and CVD were risk factors for COPD, which is identical with discoveries of earlier researches.<sup>22,35</sup> Previous studies have generally reported that CRP is an important serological marker for cardiovascular and cerebrovascular diseases. After adjusting for the effects of collinearity and covariates, we observed increased serum CRP levels to be a risk factor for COPD, suggesting that CRP as a biomarker may reflect the systemic inflammation characteristic of COPD. Previous studies have pointed out that CD8<sup>+</sup> T cells, macrophages, neutrophils, eosinophils, and epithelial cells are the main cellular components involved in COPD-related inflammation.<sup>36–38</sup> They can damage airway structures and incite systemic inflammation by releasing inflammatory elements such as IL-1, IL-6, TNF- $\alpha$ , IL-8, and CRP.<sup>39</sup> Therefore, COPD is often accompanied by complications, such as cardiovascular disease, diabetes mellitus, osteoporosis, anxiety, and depression.<sup>35</sup> However, differences in the serum levels of complement C3 and C4 between COPD patients and controls were not discovered in this research, and further mechanistic researches are needed to investigate the systemic inflammatory effect of CRP on COPD. No marked difference in immunological antibody levels between COPD

cases and controls was found in our research, which is not completely identical with earlier studies and should be supported by a bigger sample capacity and further epidemiological cohort researches. Our findings are relevant to many developing countries, where the burden of COPD is growing. Early identification of risk factors can help with prevention before disease progression and reduce the risk of disability after disease diagnosis.

Using PCA and logistic regression, we found five principal component analyses, namely PC2, PC5, PC6, PC7 and PC8, including age, education, smoking status, indoor fuel use, outdoor straw burning, RD, chd/chf, CVD and CRP, which were found to be important in COPD. As shown in our study, higher age, less education, higher smoking index (>200), exposure to outdoor straw burning, use of coal, wood, and straw indoors, history of RD, CHD, CHF, and CVD, and higher serum levels of CRP increased the prevalence of COPD. Logistic regression analysis only found that a smoking index >200, history of respiratory disease, indoor fuel use, outdoor straw burning, and higher serum levels of CRP, C3, and C3<sup>2</sup> were risk/protective factors for COPD. However, other important risk factors for COPD identified in previous studies were not significant. The 95% confidence interval of C3 was too broad; the results were not ideal, and there may have been errors. This may have been due to the collinearity among the independent variables, causing the regression equation to become unstable and resulting in the significant variables becoming insignificant. Principal component logistic regression analysis found more important risk factors for COPD, especially some widely recognized factors, which are closer to those of previous studies. Therefore, to address the problem of collinearity in this study and to identify the risk factors for COPD, PCA logistic regression analysis has more advantages than the logistic regression method.

This study also has a number of limitations. Our original contributions were not found in post-bronchodilator spirometry. Nevertheless, previous studies have indicated that only 9% of people with high-risk COPD can reverse airway obstruction after using bronchodilators, and the findings are robust.<sup>40,41</sup> In addition, this study is cross-sectional, and identification of the causal effect of the serological index on COPD requires further prospective cohort and mechanistic studies.

## Conclusion

In summary, COPD was significantly more common in people over 40 years of age in the rural Northeast areas of China with smoking index >200, exposure to outdoor straw burning, and the use of coal, wood, and straw indoors. It is recommended that local health care institutions could use the cut-off of smoking index as an indicator of smoking control and as part of interventions for basic primary health care when conducting health education and guidance for the prevention of COPD. Public policy should be taken to prevent rural residents from exposure to indoor smoke and outdoor straw burning. In addition, medical history and higher serum level of CRP are powerful predictors of COPD. Smoking cessation services and healthy lifestyle guidance should be provided for people with history of respiratory disease, cardiovascular and cerebrovascular disease, and people with higher serum level of CRP. Compared to the logistic regression analysis, PCA logistic regression analysis identified more important risk factors for COPD.

## Data Sharing Statement

The primary contributions in this study are presented in the article and Supplementary Appendix. Contact the corresponding author directly for further inquiries.

## Ethics Statement

The study was in accordance with the Declaration of Helsinki and was conducted after obtaining approval from the Ethics Committee of Harbin Medical University. Each patient signed a written informed consent.

## Funding

This study was funded by National Health Commission of the People's Republic of China and supported by central transfer payments endemic disease project. This work was also funded by Special development funds of local colleges from the central government (Study on prevention and control of major diseases in Heilongjiang Province).



## Disclosure

The authors state no conflicts of interest in this work.

## References

1. GOLD Scientific Committee, Pauwels RA, Buist AS, Calverley PM, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO global initiative for Chronic Obstructive Lung Disease (GOLD) workshop summary. *Am J Respir Crit Care Med.* 2001;163(5):1256–1276. doi:10.1164/ajrccm.163.5.2101039.
2. Lareau SC, Fahy B, Meek P, Wang A. Chronic Obstructive Pulmonary Disease (COPD). *Am J Respir Crit Care Med.* 2019;199(1):P1–P2. doi:10.1164/rccm.1991P1
3. Decramer M, Janssens W, Miravittles M. Chronic obstructive pulmonary disease. *Lancet.* 2012;379(9823):1341–1351. doi:10.1016/S0140-6736(11)60968-9
4. Rabe KF, Hurd S, Anzueto A, et al.; Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: gOLD executive summary. *Am J Respir Crit Care Med.* 2007;176(6):532–555. doi:10.1164/rccm.200703-456SO
5. Murgia N, Gambelunghe A. Occupational COPD-The most under-recognized occupational lung disease? *Respirology.* 2022;27(6):399–410. doi:10.1111/resp.14272
6. Soriano JB, Kendrick PJ, Paulson KR; GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med.* 2020;8(6):585–596. doi:10.1016/S2213-2600(20)30105-3
7. Burney P, Patel J, Minelli C, et al. Prevalence and population attributable risk for chronic airflow obstruction in a large multinational study. *Am J Respir Crit Care Med.* 2020;203(11):1353–1365.
8. World Health Statistics. Monitoring health for the SDGs. World Health Organization; 2017. Available from: [http://www.who.int/gho/publications/world\\_health\\_statistics/2017/en/](http://www.who.int/gho/publications/world_health_statistics/2017/en/). Accessed October 6, 2017.
9. IEA. *World Energy Outlook 2017*. International energy agency publications; 2017.
10. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet.* 2017;390(10100):1151–1210. doi:10.1016/S0140-6736(17)32152-9
11. Fang L, Gao P, Bao H, et al. Chronic obstructive pulmonary disease in China: a nationwide prevalence study. *Lancet Respir Med.* 2018;6(6):421–430. doi:10.1016/S2213-2600(18)30103-6
12. Sainani KL. Logistic regression. *PM R.* 2014;6(12):1157–1162. doi:10.1016/j.pmrj.2014.10.006
13. Huang T, Li J, Zhang W. Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism. *BMC Med Res Methodol.* 2020;20(1):99. doi:10.1186/s12874-020-00989-x
14. Cadima J, Cerdeira JO, Minhoto M. Computational aspects of algorithms for variable selection in the context of principal components. *Comput Stat Data Anal.* 2004;47(2):225–236. doi:10.1016/j.csda.2003.11.001
15. Ricciardi C, Valente AS, Edmund K, et al. Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health Informatics J.* 2020;26(3):2181–2192. doi:10.1177/1460458219899210
16. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans a Math Phys Eng Sci.* 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202
17. Ringner M. What is principal component analysis? *Nat Biotechnol.* 2008;26(3):303–304. doi:10.1038/nbt0308-303
18. Halbert RJ, Natoli JL, Gano A, et al. Global burden of COPD: systematic review and meta-analysis. *Eur Respir J.* 2006;28(3):523–532. doi:10.1183/09031936.06.00124605
19. Zhu B, Wang Y, Ming J, et al. Disease burden of COPD in China: a systematic review. *Int J Chron Obstruct Pulmon Dis.* 2018;27(13):1353–1364. doi:10.2147/COPD.S161555
20. Zhong N, Wang C, Yao W, et al. Prevalence of chronic obstructive pulmonary disease in China: a large, population-based survey. *Am J Respir Crit Care Med.* 2007;176(8):753–760. doi:10.1164/rccm.200612-1749OC
21. Buist AS, McBurnie MA, Vollmer WM, et al. International variation in the prevalence of COPD (the BOLD study): a population-based prevalence study. *Lancet.* 2007;370(9589):741–750. doi:10.1016/S0140-6736(07)61377-4
22. Han R, Zou J, Shen X, et al. [The risk factors of chronic obstructive pulmonary disease in Heilongjiang province]. *Zhonghua Jie He He Hu Xi Za Zhi.* 2015;38(2):93–98. Chinese.
23. Brandsma C-A, de Vries M, Costa R, et al. Lung ageing and COPD: is there a role for ageing in abnormal tissue repair? *Eur Respir Rev.* 2017;26(146):170073. doi:10.1183/16000617.0073-2017
24. D’Ascanio M, Viccaro F, Calabrò N, et al. Assessing static lung hyperinflation by whole-body plethysmography, helium dilution, and Impulse Oscillometry System (IOS) in patients with COPD. *Int J Chron Obstruct Pulmon Dis.* 2020;15(15):2583–2589. doi:10.2147/COPD.S264261
25. Meiners S, Eickelberg O, Königshoff M. Hallmarks of the ageing lung. *Eur Respir J.* 2015;45(3):807–827. doi:10.1183/09031936.00186914
26. Sobrino E, Irazola VE, Gutierrez L, et al. Estimating prevalence of chronic obstructive pulmonary disease in the Southern Cone of Latin America: how different spirometric criteria may affect disease burden and health policies. *BMC Pulm Med.* 2017;17(1):187. doi:10.1186/s12890-017-0537-9
27. He MZ, Zeng X, Zhang K, Kinney PL. Fine particulate matter concentrations in urban Chinese cities, 2005–2016: systematic review. *Int J Environ Res Public Health.* 2017;14(2):191. doi:10.3390/ijerph14020191
28. Cheng X, Li J, Zhang Z. Analysis of basic data of the study on prevention and treatment of COPD and chronic cor pulmonale. *Zhonghua Jie He He Hu Xi Za Zhi.* 1998;21:749–752.
29. Xu F, Yin X, Zhang M, Shen H, Lu L, Xu Y. Prevalence of physician-diagnosed COPD and its association with smoking among urban and rural residents in regional mainland China. *Chest.* 2005;128:2818–2823.
30. Gershon AS, Warner L, Cascagnette P, et al. Lifetime risk of developing chronic obstructive pulmonary disease: a longitudinal population study. *Lancet.* 2011;378(9795):991–996. doi:10.1016/S0140-6736(11)60990-2

31. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet*. 2007;370(9589):765–773. doi:10.1016/S0140-6736(07)61380-4
32. Liu S, Zhou Y, Wang X, et al. Biomass fuels are the probable risk factor for chronic obstructive pulmonary disease in rural South China. *Thorax*. 2007;62(10):889–897. doi:10.1136/thx.2006.061457
33. Hopkinson NS. COPD, smoking, and social justice. *Lancet Respir Med*. 2022;10(5):428–430. doi:10.1016/S2213-2600(22)00130-8
34. Pezzuto A, Lionetto L, Ricci A, et al. Inter-individual variation in CYP2A6 activity and chronic obstructive pulmonary disease in smokers: perspectives for an early predictive marker. *Biochim Biophys Acta Mol Basis Dis*. 2021;1867(1):165990. doi:10.1016/j.bbadis.2020.165990
35. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for diagnosis, management, and prevention of COPD; 2021. Available from <http://www.goldcopd.org/>. Accessed November 20, 2021.
36. Quint JK, Wedzicha JA. The neutrophil in chronic obstructive pulmonary disease. *J Allergy Clin Immunol*. 2007;119(5):1065–1071. doi:10.1016/j.jaci.2006.12.640
37. Gamble E, Grootendorst DC, Hattotuwa K, et al. Airway mucosal inflammation in COPD is similar in smokers and ex-smokers: a pooled analysis. *Eur Respir J*. 2007;30(3):467–471. doi:10.1183/09031936.00013006
38. Baraldo S, Turato G, Badin C, et al. Neutrophilic infiltration within the airway smooth muscle in patients with COPD. *Thorax*. 2004;59(4):308–312. doi:10.1136/thx.2003.012146
39. Tanni SE, Pelegrino NR, Angeleli AY, et al. Smoking status and tumor necrosis factor-alpha mediated systemic inflammation in COPD patients. *J Inflamm*. 2010;7(1):29. doi:10.1186/1476-9255-7-29
40. Johannessen A, Omenaas ER, Bakke PS, Gulsvik A. Implications of reversibility testing on prevalence and risk factors for chronic obstructive pulmonary disease: a community study. *Thorax*. 2005;60(10):842–847. doi:10.1136/thx.2005.043943
41. Dement J, Welch L, Ringen K, et al. A case-control study of airways obstruction among construction workers. *Am J Ind Med*. 2015;58(10):1083–1097. doi:10.1002/ajim.22495

## Risk Management and Healthcare Policy

Dovepress

### Publish your work in this journal

Risk Management and Healthcare Policy is an international, peer-reviewed, open access journal focusing on all aspects of public health, policy, and preventative measures to promote good health and improve morbidity and mortality in the population. The journal welcomes submitted papers covering original research, basic science, clinical & epidemiological studies, reviews and evaluations, guidelines, expert opinion and commentary, case reports and extended reports. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/risk-management-and-healthcare-policy-journal>