Risk assessments of chronic obstructive pulmonary disease using Bayesian network based on a provincial survey

Authors: Changyue Shangguan, Lianzheng Yu, Guangcong Liu, Yibing Song, Jie Chen

Article type: Original article

Received: November 25, 2020.

Accepted: March 5, 2021.

Published online: March 5, 2021.

ISSN: 1897-9483

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), allowing third parties to copy and redistribute the material in any medium or format and to remix, transform, and build upon the material, provided the original work is properly cited, distributed under the same license, and used for noncommercial purposes only. For commercial use, please contact the journal office at pamw@mp.pl.
Risk assessments of chronic obstructive pulmonary disease using Bayesian network based on a provincial survey

Short title: A Model to Assess Chronic Obstructive Pulmonary Disease Risks

Changyue Shangguan¹, Lianzheng Yu ², Guangcong Liu³, Yibing Song¹, Jie Chen¹*

Departmental and institutional affiliation:

1 Department of Occupational and Environmental Health, School of Public Health, China Medical University, Shenyang, People's Republic of China

2 Department of Noncommunicable Chronic Disease Prevention, Liaoning Center for Disease Control and Prevention, Shenyang, People's Republic of China.

3 Department of Epidemiology, Cancer Hospital of China Medical University, Liaoning Cancer Hospital and Institute Shenyang, Shenyang, People's Republic of China

Corresponding author:

Jie Chen, M.D. (Institutional phone number: +86-024-31939406)

Department of Occupational and Environmental Health, School of Public Health, China Medical University, Shenyang, People's Republic of China.

E-mail: jchen@cmu.edu.cn.

Conflict of Interest Statement: none declared
What's new?

We carried out a provincial study investigating the Chronic Obstructive Pulmonary Diseases (COPD) prevalence and its risk factors in Liaoning Province, China. In addition to the previously identified risk factors, we found lower air quality satisfaction was also the potential risk factor for COPD. We built a Bayesian network to assess the individual COPD risks, which was based on questionnaires. The model had a good performed with a mean area under the curve (AUC) of 0.85 and the optimum accuracy of 0.87 (cutoff=0.473). From this model, we also found that the risk of a combination of several risk factors was considerably larger than the risk of any single risk factor, and such kind of population deserved more attention in COPD prevention.

**Key words:** Bayesian Networks, Chronic Obstructive Pulmonary Disease, Risk assessment, Screening
Abstract

**Introduction:** The diagnosis of chronic obstructive pulmonary disease (COPD) is based on spirometry tests that are difficult to perform in some populations.

**Objectives:** We aimed to construct a risk assessment model using a Bayesian Network (BN) that would enable screening high-risk populations.

**Patients and methods:** A provincial survey of COPD was performed with face-to-face interviews and spirometry tests among the population aged ≥40 years in Liaoning Province, northeastern China. The potential risk factors were initially identified by multivariable logistic regression, and then a BN was built. To validate its performance, cross-validation and external dataset validation were performed, and area under the curve (AUC) and accuracy of the BN were calculated.

**Results:** The estimated age-adjusted prevalence of COPD in the entire population was 21.23% (95% confidence interval [CI]: 18.35%–24.11%). The logistic regression revealed that low education level (OR=2.35, p<0.001), elderly age (OR=4.19, p<0.001), ever smoking (OR=1.49, p=0.03), lower air quality satisfaction (OR=1.55, p=0.03) were associated with COPD. For the BN, frequent cough was the strongest single risk indicator of COPD (risk=0.374). The risks increased as more factors were specified, and the top risk was 0.738, which included the combination of elderly age, smoking, wheezing during sickness, and frequent cough. The cross-validation indicated that BN performed better than logistic regression, with a mean AUC of 0.85 and the optimum accuracy of 0.87 (cutoff=0.473).

**Conclusions:** The BN had a favorable performance in predicting COPD risks based on questionnaires. The risks associated with the combination of several risk factors should be noted.
Introduction

Chronic obstructive pulmonary disease (COPD) is one of the most common causes of respiratory disease around the world: there were 251 million cases of COPD globally in 2016. COPD also accounted for more than 5% of all causes of mortality in the previous year[1]. Most deaths occurred in developing countries, and the prevalence differed among regions[2]. For example, Muñozes et al[3] reported a range of COPD prevalence, with the lowest in Mexico City (7.8%) and the highest in Montevideo, Uruguay (19.7%). Similarly, according to some studies, COPD prevalence varies among different regions in China, ranging from 2% to 21%, and is generally higher in the west than in the east[4-7]. These variations across geographic areas are attributed partly to the different levels of risk exposure such as tobacco use rate, age structure, and disparities in socio-economic development, including income and healthcare level[2]. However, it is critical to identify risk factors in the management of COPD. Numerous studies of COPD prevalence have shown that male sex, older age, low education level, and tobacco use are associated with the development of COPD[7-12], but the risks identified in the studies performed in different regions varied. Therefore, a regional survey of COPD, as well as its risk factors, is still needed.

Many nations spend considerable budgets on COPD[13-17]. Early diagnosis is one of the best ways to help save costs, but the time-consuming pre-bronchodilator and post-bronchodilator spirometry, whose results are the main basis for the diagnosis of COPD, cannot be performed in all populations, especially in middle- and low-income states. Still, screening individuals at high risk for COPD using spirometry tests is necessary. In previous studies, multivariable logistic regressions were frequently used to identify the risk factors for COPD, but they are not often used in risk assessments or predictions. The stability of the logistic regressions is largely affected by the potential collinearity among the included
variables, and the predictions of continuous variables are not available[18, 19]. Fortunately, machine learning strategies offer an alternative approach. As a powerful and increasingly popular tool for diagnosis and risk assessments in biomedical fields, Bayesian Networks (BNs) can infer the probability of unknown nodes via the status of known nodes, and they can handle situations of uncertainties[20-23]. Several studies have indicated the superiority of BNs in prediction compared to logistic regression[24-27].

This study aimed to explore a risk assessment model for COPD based on a provincial cross-sectional survey of COPD. Through a BN, we can assess the risk of COPD of the individuals and then predict their disease status. If this model performs well, it will provide strong evidence for the allocation of spirometry and improve and achieve early diagnosis and treatment, which would significantly improve the cost-effectiveness of COPD management.

Patients and methods

Epidemiological survey

We conducted a cross-sectional study in Liaoning Province, northeastern China. It was launched by the National Health Commission of China and consisted of face-to-face questionnaire interviews and standardized spirometry tests. Using a multi-stage sampling strategy, 4 representative counties/districts of 4 different cities in Liaoning Province were selected by population proportion, geographical location, and economic development. Within each selected county (district), 2 adjacent communities or 2 administrative villages were randomly chosen, and at least 100 households of villagers were randomly selected from each community or administrative village. In the next step, one family member who was ≥40 years old was chosen randomly from each household using a Kish selection table. We attempted to contact each potential sample with an in-person visit. Those who did not respond after 3
contact attempts were considered ineligible, and then they were replaced with another household of similar characteristics from the same village.

For the interviews, we constructed a comprehensive questionnaire to collect information including demographic characteristics, medical history, smoking status, indoor pollution, and occupational exposure. All staff were trained to conduct the interviews before the survey began. Individuals who had resided in the selected regions for ≥6 months and were aged ≥40 years old were considered eligible for the investigation. We excluded those who lived in a communal residence; those who had cognitive, language, or mental disorders (in consideration of the interview); those with cancer (including both newly diagnosed cases and those under treatment); those with paraplegia; pregnant or breastfeeding women; and those who did not provide written informed consent. This survey was approved by the ethics review committee of the National Center for Chronic and Noncommunicable Disease Control and Prevention of the Chinese Center for Disease Control and Prevention. All included participants provided written informed consent.

**Spirometry**

The operators of the spirometry assessments, who underwent a specific training course before conducting the tests, were from local clinics in the participants’ residential areas. COPD was diagnosed according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) lung function criteria (forced expiratory volume in 1 second [FEV1]/forced vital capacity [FVC] < 0.7)[28]. All included participants underwent spirometry tests with the same brand of spirometer (MasterScreen Pneumo, Jaeger, Germany) by trained staff, following standard guidelines of the American Thoracic Society[28, 29]. First, all eligible participants underwent pre-bronchodilator spirometry. Then, post-bronchodilator spirometry was performed 15 minutes after inhalation of 400 µg salbutamol (Ventolin; GlaxoSmithKline,
Middlesex, UK). The participants who were allergic to salbutamol or those who had a resting heart rate of more than 100 beats per minute were excluded from the post-bronchodilator testing. We recorded the FEV1 and FVC during spirometry and assessed their severity of COPD with the GOLD 2017 ABCD assessment tool. Then we assessed their respiratory symptoms using the modified Medical Research Council dyspnea score. To exclude participants with other recent cardiopulmonary diseases, we also offered additional chest radiography to individuals whose post-bronchodilator FEV1/FVC was less than 70%. Using a quality grade (A to F) based on the acceptable maneuvers and the repeatability of FEV1 and FVC, the quality of each participant’s spirometry result was assessed within 24 hours. Grades A, B, and C were considered acceptable, and the others were excluded.

**Statistical methods and BNs**

Data were collected and checked by the Liaoning Center for Disease Control and Prevention. We calculated COPD standardized prevalence using sampling weights and age distribution of the aged 40 years or older. Then, multivariable logistic regression was performed using the final dataset to identify the potential risks, as well as to provide clues for the variables that would form the BN model. Since there were too many variables derived from the questionnaire, only the critical variables were entered in the logistic regression, as well as in the BN prediction model. Therefore, a feature selection was done before the logistic regression. During this procedure, the variables that contained similar or duplicated information were left out except the one that contained the most information. Then, we excluded the variables that were seldom reported to be correlated with COPD, and variables with too many null values (>50%) were excluded. The indicators from the spirometry, such as FEV1, were not included in BN because they were variables for the diagnosis of COPD. To make the logistic regression more comparable to the BN, we transformed continuous variables into categorical variables. Variables with generally acknowledged standards (such as body
mass index [BMI] or hypertension) were categorized according to these standards. Otherwise, the variables (such as income) were categorized into 3 groups (high, medium, and low). The remaining variables were included in the logistic regression. All tests were 2-sided, and a p-value of less than 0.05 was considered statistically significant.

BN (also known as Bayesian Belief Network), is a directed acyclic graph (DAG) that represents nodes (variables) and their conditional probabilities. In this study, we used BN to construct a risk assessment model for COPD. The variables of our BN were based on the features identified in logistic regression and were supplemented by the current evidence.

The learning of the BN included structure learning and parameter learning[30]. For the structure learning, the Hill Climbing (HC) Algorithm was adopted. As to the a priori of the basic structure, the established causal relationships between some variables were acknowledged. For example, older age, occupational exposure to dust, and ever smoking have been reported to be associated with a higher risk of COPD. These causal relationships were defined as the “whitelist”. On the other hand, some seemingly impossible relationships between the variables were also known. For instance, none of the characteristics can affect age or sex, and these implausible relationships were defined as the “blacklist”. With the help of the “bnlearn” package of R statistics, we achieved such HC learning based on setting the whitelist and blacklist[31]. To exert the flexibility of BN in dealing with uncertainties, we tried to set the least a priori as possible. For the parameter learning, we chose the Bayes method rather than the maximum likelihood estimation (MLE) method because its estimated parameters are smoother, making the inference both easier and more robust[30].

We transferred the continuous variables into the categorical to better fit the BN. For some continuous variables, we divided them into 3 groups according to tertiles. But for some variables, specific cutpoints were used. For example, the variable "Age" was divided into 3
groups: young age group (younger than 50-year-old), middle-age group (50~59-year-old), and older-age group (elder than 59-year-old). And the participants were defined as hypertension according to the blood pressure if his/her systolic blood pressure was higher than 140mmHg, or the diastolic blood pressure was higher than 90 mmHg. The definition of high, middle, and low education level was >9, 6~9, and <6 years respectively.

In this study, the conditional probabilities of COPD under specific circumstances were adopted as indicators of risk. Note that these conditional probabilities generated by “bnlearn” were generated using Monte Carlo particle filters, so each run of the prediction may yield slightly different values. Firstly, we divided the dataset into training-set (containing 90% of the observations) and test-set (containing 10% of the observations) using random sampling. The BN was constructed based on the training set. To make the prediction more robust, we ran each prediction 100 times and took the median as the final prediction. Then, we calculated the conditional probabilities under all the circumstances that specified 1 factor, 2 factors, 3 factors, and 4 factors, and then we selected the top 20 cases for each group. To validate the BN, both internal and external validations were adopted. For the internal validation, we carried out 5-fold cross-validation for both the BN and the logistic regression, during which we depicted the receiver operating characteristic (ROC) curves and calculated the areas under the curves (AUCs) and accuracy. For the external validation, the BN was validated using the test-set. We considered an acceptable AUC of 0.75, and BNs with lower AUC were reconstructed until they met the threshold. The research process of this study is shown in Figure 1.

All statistical analyses and BN-related procedures were completed with R statistics version 3.6.2, from The Comprehensive R Archive Network (http://cran.r-project.org/).
Results

Demographic characteristics

Between December 2014 and December 2015, a total of 2400 participants from 4 counties/districts of 4 cities (Xinmin County of Shenyang City, Mingshan District of Benxi City, Donggang County of Dandong City, and Haizhou District of Fuxin City) were selected. Among these, 2397 (99.88%) were qualified and interviewed, while the other 3 (0.02%) were excluded for failing to meet the interviewing criteria. Of all the interviewed respondents, 167 individuals were ineligible for spirometry, and the other 2230 completed spirometry, which included pre-bronchodilator and post-bronchodilator examinations. Therefore, a total of 2194 participants completed the whole procedure and were included in the final analyses. Details are shown in Figure S1.

Figure S2 and Figure S3 show the demographic characteristics of the 2194 participants, among whom 1205 were women (54.9%) and 989 were men (45.1%). Roughly half of the participants (n=1107 [50.5%]) were from rural areas and the other half (n=1087 [49.5%]) were from urban areas. The proportion of participants who received a higher level of education (educated >9 years) was 27.1% (n=594). The numbers of participants who were ever exposed to occupational hazardous gas and dust were 692 (31.5%) and 778 (35.5%), respectively. Two-thirds of the participants (n=1461 [66%]) suffered from hypertension, and 823 (37.5%) were smokers, of which 689 (83.72%) were men and only 134 (16.28%) were women. The overweight and obesity rates among the participants were 2.3% and 26.8%, respectively. According to spirometry and further chest radiography examination, the age-adjusted prevalence of COPD in adults aged ≥40 years was 21.23% (95% confidence interval [CI] 18.35%–24.11%), in which the age-adjusted prevalence among men (23.89%, 95% CI:
16.72%–31.05%) was significantly higher than that among women (18.88%, 95% CI: 16.66%–21.11%).

**Multivariable logistic regression**

A total of 101 variables from the questionnaire were included through primary selection. After further examining the relationships with COPD, comparing similar variables, as well as excluding those with too many null values, 37 variables were selected. The observations with null values were also excluded. Finally, 1656 observations were entered in the multivariable logistic regression. It revealed that female sex (odds ratio [OR]=0.66, p=0.02) was the protective factors for COPD, while other factors such as low and middle education level (OR=2.35, p<0.001 and OR=1.74, p=0.004), diagnosed coronary heart disease (CHD) (OR=1.72, p=0.03), parental emphysema history (OR=2.59, p<0.001), ever smoking (OR=1.49, p=0.03), older age (OR=4.19, p<0.001) and middle age (OR=2.09, p<0.001), and middle level of income (OR=1.45, p=0.02) were identified as risks for COPD. Interestingly, we also found that the less satisfied participants were with current air quality, the higher the risk for COPD (OR=1.72, p=0.03). Contrary to many previous studies, childhood environmental tobacco smoke exposure was found to be negatively associated with COPD. It was difficult to explain based on this dataset, but we found the proportion of pulmonary heart disease in the non-exposure group is higher than that in the exposure group (1.91% versus 0.88%). Additionally, the COPD prevalence (unadjusted) was also higher in the non-exposure group (18.53% versus 16.47%). Details of the results of the logistic regression are presented in Table 1.

**Bayesian Network**

All significant risk or protective factors were entered in the BN. Additionally, the variables that were frequently reported to be associated with COPD, such as BMI[32-34],
occupational exposure to dust or hazardous gas, exposure to coal or biomass fuel, and rural residence[10, 35, 36], were also considered. Therefore, a total of 17 features (including COPD) were finally selected. In the next step of setting a priori, for the whitelist, age, smoking status, BMI, and sex were all set to be associated with COPD. For the blacklist, age, sex, and parental emphysema were not influenced by most of the other variables. The a priori containing the whitelist and the blacklist is shown in Table S1 and Table S2 in the supplementary files.

The structure of the BN is shown in Figure 2. Apart from the a priori that we set, no other parent or child associations with COPD were found. However, some relationships between other nodes were found by structure learning. For example, sex (the node “Sex”) affected ever smoking (the node “Smoke”), and ever smoking also led to frequent cough (the node “Cough”); education level was the parent of income level (the node “Income”) and residence area (the node “Region”). The residence area also affected occupational exposure to hazardous gas (the node “Gas”) or dust (the node “Dust”), which might be attributed to regional characteristics, and CHD (the node “CHD”) was affected by age. Most of the structure learning-based relationships agree with the current evidence, implying the plausible validity of this BN.

Part of the parameter learning results is shown in Figure 3. Figure 3 represents the top 20 high-risk circumstances when specifying 1 factor (A), 2 factors (B), 3 factors (C), and 4 factors (D). In the circumstances specifying a single factor, the highest risk was frequent cough (“Cough” in Figure 3, risk=0.374): specifically, the risk of an individual with frequent cough (in the population aged ≥40 years, which matches our sampling population) was 0.374. Wheezing during sickness (WSS in Figure 3, risk=0.326) and the elderly age (Elder in Figure 3, risk=0.293) ranked second and third, respectively. The top 20 risks kept increasing when more factors were specified. When 2 factors were specified, the highest risk increased to
0.553, which corresponded to the combination of wheezing during sickness and frequent cough. The combination of elderly age and frequent cough had a risk of 0.511, and the combination of elderly age and wheezing during sickness had a risk of 0.453. The top risk when specifying 3 factors was the combination of elderly age, wheezing during sick, and frequent cough (risk=0.681), and the top risk when specifying 4 factors was the combination of elderly age, ever smoking, wheezing during sickness, and frequent cough (risk=0.738).

We found that the risks of elderly age, frequent cough, ever smoking, wheezing during sickness generally ranked at the forefront when specifying single factors. However, when an increasing number of factors was specified, the role of middle income, lower education level, and obesity tended to play more important roles. For example, the risk of COPD when specifying middle income was only 0.224, but the risk of the combination of wheezing during sickness, frequent cough, and middle income increased to 0.605, which ranked eighth in Figure 3(C), and the risk of the combination of elderly age, frequent cough, wheezing during sickness, and middle-income level was 0.721, which ranked seventh in Figure 3 (D).

The cross-validation indicated that the mean AUC of the BN was 0.85, with a mean accuracy of 0.87 (cutoff point=0.473). The mean AUC and accuracy of the logistic regression were 0.77 and 0.84, respectively. For the external validation (shown in Figure S4), the AUC and accuracy were 0.82 and 0.86 (cutoff=0.40), respectively. These suggested that the BN had a favorable performance for predicting the risk of COPD, which is also much better than the logistic regression model. The ROC curves created during the cross-validations and external validation are presented in Figure 4/Table S3 and Figure S4, respectively.

**Discussion**

This study is the first official population-based survey of COPD in northeastern China in the past decade. We found the age-adjusted COPD prevalence was 21.23% among people
aged ≥40 years, which was considerably higher than the reported regional prevalence[4, 6] and nationwide prevalence (13.6%)[33]. This prevalence was also higher than most regional studies with similar designs around the world[3, 37, 38], except for Maastricht in the Netherlands (24.0%)[38]. The main reason for this may be the differences in risk exposure. For example, the proportions of the population exposed to coal or biomass fuels and occupational dust were 38.7% and 36.5%, respectively; in most regions, however, the estimated proportions of exposure to the same risks ranged from 0.90% to 37.5% and 16.60% to 27.34%, respectively[3, 39-42].

In the logistic regression, not only did we identify age, sex, education level, and ever smoking as influencing factors of COPD, but we also named middle-income level and lower levels of air quality satisfaction as risk factors. Middle-income has seldom been reported as a risk for COPD, but, in this study, we defined middle-income level as an annual income of 20,000–48,000 CNY (about $2900 to $6900), which was a relatively low-income level. Additionally, one of the main symptoms of COPD is shortness of breath, which would increase the vulnerability to bad air quality. Therefore, the lower level of air satisfaction may reflect COPD status. Interestingly, Childhood environmental tobacco exposure was found to be a protective factor, and cooker hood (compared to no hood) was a risk factor. The reason for this was difficult to explain, since the proportions of other risk factors in these two groups were higher than or equal to the proportions of no childhood exposure group or no hood group, respectively. However, two recent studies carried out in China reported similar results. Li et al. indicated hood as a risk factor in his study, which was carried out in the same provinces as our study[43]. Another national study in China failed to found the association between smokings living in the home and COPD[44]. Therefore we suggest these findings might not be caused by bias, however, the reasons for this need further exploration.
The conditional probabilities generated by the BN provide new perspectives in predicting COPD among the community population. Given that the diagnosis of COPD requires pre-bronchodilator and post-bronchodilator spirometry, it is of great benefit to find a simple way of screening the population that is at high risk. Previously, logistic regressions in healthcare-related research mostly focused on identifying risks and providing evidence for disease preventions and treatments. Different from that approach, BNs calculate conditional probabilities (which we used as the risks of diseases) under a variety of situations. This is more practical for personalized predictions since it is common that the population would be exposed to several risks at the same time. We also found that the risks of the individuals who were exposed to 4 risk factors were significantly higher than those individuals exposed to 1 or 2 factors. These findings imply that although some of these involved factors were not the main influencing factors (or indicators) of COPD risks, they may exert significant effects under some specific circumstances. Therefore, individuals exposed to a combination of risk factors deserve more attention.

Figure 4 indicates that the performance of the BN was better than the logistic regression during the cross-validation, and the external validation confirmed the performance of the BN. Moreover, BNs also identify interactions of all included nodes, as well as the adjacent nodes of the targets. So, if we can further confirm the risk factors for COPD and their stages in the network of the disease process, we will gain more insights into certain high-risk behaviors through BNs, and more targeted measures can be taken to better prevent COPD.

BNs can reflect relationships between all adjacent nodes, which also helps provide evidence for exploring further interactions between the target and each factor. All the reasonable relationships between the nodes mentioned in the “Results” section were identified by machine learning, implying the strong causal inference of this COPD-related BN prediction model. However, there are also some uncommon causal relationships: residence
area was impacted by sex and education level and air satisfaction (the node “Air_sat”) was impacted by ventilation of the kitchen (the node “Venti”). Still, these interactions can be explained. First, more women were sampled in the urban areas, which resulted in the interactions between sex and residence area. Second, imbalances in development between rural and urban areas are larger in developing countries like China. Therefore, the more educated population tends to reside in urban areas for better jobs and public services, which led to the relationship between education level and residence area. Third, the education level can affect not only the residence but also the perception of health, including the ventilation of the kitchens. The proportions of people with high, medium, and low education levels who only used chimneys without ventilation equipment were 30.8%, 34.3%, and 39.3%, respectively. Further, those who were exposed to coal or biomass fuel when cooking suffered from heavy indoor air pollution and this may lead to a higher demand for ambient air quality.

Excellent prediction models for COPD are of considerable significance for public health. However, the diagnostic standard of COPD requires time-exhausting pre-bronchodilator and post-bronchodilator spirometry, which is difficult to perform in the total population. Prediction models, such as the BN in this study, can help initially assess the risks of individuals based only on an interview or questionnaire and screen the high-risk population for spirometry for further examination. In this way, if we can identify most COPD patients in the early stages of the disease, timely interventions will significantly prevent them from disease progression. In addition, other tools such as the COPD assessment test (CAT) questionnaires were also proved to be efficient in evaluating the COPD patients[45-48]. Therefore, the combined applications of these prediction related tools should be recommended. With their help, severe COPD cases could be decreased, which also helps save COPD-related spending and improves cost-efficiency. Therefore, such models are particularly helpful in developing countries with large populations and limited healthcare resources.
The BN we constructed was validated to perform favorably in assessing and predicting COPD risks, which confirmed the contribution of this study. Nevertheless, some limitations must be noted. First, the data of our BN was from a cross-sectional survey rather than a cohort study: cross-sectional designs are limited in their ability to confirm risk factors for COPD. Second, all variables were collected from face-to-face interviews: recall bias could not be avoided despite our strict quality control. Third, the performance of our BN still needs to be improved, although the current AUC and accuracy are acceptable. One possible way to improve the BN is to include more critical variables, which is another limitation in this study. In the future, prediction models for COPD based on high-quality cohort and case-control studies are needed.

Conclusions

The prevalence of COPD in the population of Liaoning aged ≥40 years was 21.32%, which is higher than in most regions. Our BN had a favorable performance in predicting the risks of COPD which was constructed based on questionnaires, and individuals exposed to a combination of risk factors deserve increased attention.

Contribution statement

CY Shangguan and J Chen conceived of the study and participated in its design and coordination. CY Shangguan drafted the manuscript. LZ Yu and GC Liu, and Shangguan CY take responsibility for data analysis. LZ Yu, GC Liu, YB Song Shangguan contributed to critical revision. All authors read and approved the final manuscript.

Acknowledgments

This study received no funding.
References


26. Le QA, Strylewicz G, Doctor JN. Detecting blood laboratory errors using a Bayesian


### Table 1. Multivariable logistic regression

<table>
<thead>
<tr>
<th>Variables</th>
<th>Beta</th>
<th>OR</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>cooker hood (/chimney)</td>
<td>0.484</td>
<td>1.62</td>
<td>1.16~2.28</td>
<td>0.005</td>
</tr>
<tr>
<td>ventilator (/chimney)</td>
<td>0.013</td>
<td>1.01</td>
<td>0.72~1.43</td>
<td>0.94</td>
</tr>
<tr>
<td>Female (Versus Male)</td>
<td>-0.417</td>
<td>0.66</td>
<td>0.47~0.93</td>
<td>0.02</td>
</tr>
<tr>
<td>Low education level (High education level)</td>
<td>0.855</td>
<td>2.35</td>
<td>1.54~3.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Middle education level (High education level)</td>
<td>0.556</td>
<td>1.74</td>
<td>1.19~2.55</td>
<td>0.004</td>
</tr>
<tr>
<td>Low level of income (High level of income)</td>
<td>-0.149</td>
<td>0.86</td>
<td>0.56~1.32</td>
<td>0.5</td>
</tr>
<tr>
<td>Middle level of income (High level of income)</td>
<td>0.374</td>
<td>1.45</td>
<td>1.05~2.01</td>
<td>0.02</td>
</tr>
<tr>
<td>CHD unknown (Not CHD)</td>
<td>0.109</td>
<td>1.12</td>
<td>0.52~2.41</td>
<td>0.78</td>
</tr>
<tr>
<td>CHD (Not CHD)</td>
<td>0.543</td>
<td>1.72</td>
<td>1.07~2.78</td>
<td>0.03</td>
</tr>
<tr>
<td>Parental emphysema unknown (No parental emphysema)</td>
<td>-0.343</td>
<td>0.71</td>
<td>0.32~1.59</td>
<td>0.41</td>
</tr>
<tr>
<td>Parental emphysema (No parental emphysema)</td>
<td>0.95</td>
<td>2.59</td>
<td>1.63~4.11</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Wheezing sound during sick unknown (No wheezing sound during sick)</td>
<td>-0.822</td>
<td>0.44</td>
<td>0.05~4.16</td>
<td>0.47</td>
</tr>
<tr>
<td>Wheezing sound during sick (No wheezing sound during sick)</td>
<td>0.715</td>
<td>2.04</td>
<td>1.45~2.88</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Frequent coughing (No frequent coughing)</td>
<td>0.776</td>
<td>2.17</td>
<td>1.3~3.63</td>
<td>0.003</td>
</tr>
<tr>
<td>Childhood environmental tobacco exposure unknown (NO childhood environmental tobacco exposure)</td>
<td>0.993</td>
<td>2.7</td>
<td>0.58~12.65</td>
<td>0.21</td>
</tr>
<tr>
<td>Childhood environmental tobacco exposure (NO childhood environmental tobacco exposure)</td>
<td>-0.364</td>
<td>0.7</td>
<td>0.52~0.93</td>
<td>0.01</td>
</tr>
<tr>
<td>Ever smoking (Never smoking)</td>
<td>0.4</td>
<td>1.49</td>
<td>1.05~2.12</td>
<td>0.03</td>
</tr>
<tr>
<td>Elder age group (Younger age group)</td>
<td>1.432</td>
<td>4.19</td>
<td>2.79~6.29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Middle age group (Younger age group)</td>
<td>0.738</td>
<td>2.09</td>
<td>1.37~3.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>lower level of air satisfaction (Higher level of air satisfaction)</td>
<td>0.44</td>
<td>1.55</td>
<td>1.05~2.29</td>
<td>0.03</td>
</tr>
<tr>
<td>Middle level of air satisfaction (Higher level of air satisfaction)</td>
<td>-0.074</td>
<td>0.93</td>
<td>0.63~1.36</td>
<td>0.7</td>
</tr>
<tr>
<td>Tachycardia (Not Tachycardia)</td>
<td>-0.925</td>
<td>0.4</td>
<td>0.12~1.26</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Age group: younger than 50-year-old, young age group (reference); 50~59-year-old, middle age group; elder than 59-year-old, older age.

Income level: <20,000 Chinese yuan (CNY), low-income level; 20,000~48,000 CNY, middle-income level; >48,000 CNY, high income level (reference).

Abbreviations: CHD, Coronary Heart Disease.
Figure 1. Flow chart of this study. The participants were randomly sampled in 4 cities of Liaoning Province, China. The participants were interviewed, and they also underwent pre-bronchodilator and post-bronchodilator spirometry, if qualified. The results were examined by
the staff, and the qualified observations were included in the multivariate analysis. The Bayesian Network was constructed based on a priori that consisted of multivariate logistic results and previous evidence. Validations were followed to validate the model, and an AUC of at least 0.75 was considered acceptable.

![Bayesian Network Diagram](image)

**Figure 2. The structure of the Bayesian Network.** Some relationships were found between the variables by structure learning. The network reflected the interactions between all nodes, not only for Chronic Obstructive Pulmonary Disease. For example, sex affected smoking status; education level affected income level and residence area (the node “Region” in this
Figure); residence area affected occupational exposure to hazardous gas (the node “Gas”) or dust (the node “Dust”), and coronary heart disease (the node “CHD”) was affected by age. All the relationships are plausible according to current evidence. For the nodes: Age: age group; Air_sat: satisfaction level of ambient air quality; BMI: body mass index; CHD: coronary heart disease; Coalbio: exposure to coal or biomass fuel when cooking; Cough: frequent coughing; Dust: occupational exposure to dust; Edu: education level; Income: income level; Smoke: ever smoking; sm14: environmental tobacco exposure before 14; Region: resident areas; Venti: ventilation of the kitchen; Wheezing: wheezing during sickness.
Figure 3. The top high-risk conditional probabilities specifying 1(A), 2(B), 3(C), and 4(D) factors. This figure represents the top 20 high-risk circumstances when specifying 1 factor (A), 2 factors (B), 3 factors (C), and 4 factors (D). Chimney: using a chimney as ventilation in the kitchen; Cough: frequent cough; Coalbio: exposure to coal or biomass fuel during cooking; Dust_exp: occupational exposure to hazardous dust; Elder: elderly age; Gas_exp:
occupational exposure to hazardous gas; L-AS: lower level of air quality satisfaction; L_edu: lower education level; M_inc: middle-income level; SM14: environmental tobacco smoke before 14 years old; Smoke: ever smoking; WSS: wheezing sound during sickness.

Figure 4. The receiver operating characteristic curves were created during the 5-fold cross-validations. Each panel (panel A, B, C, D and E) included the ROC curves of logistic regression (the red lines) and the Bayesian Network (the blue lines) during one fold of the validation.