Exploring Related Factors of Chronic Obstructive Pulmonary Disease Based on Elastic Net and Bayesian Network

Dichen Quan
Shanxi Medical University
Jiahui Ren
Shanxi Medical University
Hao Ren
Shanxi Medical University
Liqin Linghu
Shanxi Medical University
Xuchun Wang
Shanxi Medical University
Meichen Li
Shanxi Medical University
Yuchao Qiao
Shanxi Medical University
Zeping Ren
Shanxi Centre for Disease Control and Prevention
Lixia Qiu (qlx_1126@163.com)
Shanxi Medical University

Research Article

Keywords: COPD, Elastic Net, MMHC algorithm, Bayesian Networks

Posted Date: November 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1051771/v1

Abstract

Objective

This study aimed to construct Bayesian networks to analyze the network relationship between COPD and its related factors, and to explore the influencing intensity on COPD through network reasoning.

Method

Firstly Elastic Net and MMHC hybrid algorithm were adopted to screen the variables of the data of COPD in Shanxi Province from 2014 to 2015 and construct Bayesian networks respectively, and the parameters were estimated by maximum likelihood estimation.

Results

After feature selection by Elastic Net, 10 variables closely related to COPD finally entered the model. The COPD Bayesian networks constructed by MMHC algorithm showed that smoking status, household air pollution, family history, cough, air hunger or dyspnea were directly related to COPD, in which smoking status, household air pollution and family history were the parent nodes of COPD, and cough, air hunger or dyspnea represented the child nodes of COPD. In other words, smoking status, household air pollution, family history were related to the occurrence of COPD, and COPD would affect cough, air hunger or dyspnea. Gender was indirectly linked to COPD through smoking status.

Conclusion

Using Elastic Net to knock out some weakly-associated influencing factors of COPD in the variable screening stage, Bayesian networks could reveal the complex network relationship between COPD and its relevant factors well, making it more convenient to carry out targeted prevention and control of COPD. As such, Bayesian networks enjoyed a good prospect of application in analyzing disease-related factors.

Introduction

Chronic obstructive pulmonary disease (COPD) is a common disease characterized by persistent respiratory symptoms and airflow limitation. Its clinical manifestations comprise cough, expectoration, chest tightness, air hunger and dyspnea. In severe cases, it can progress into respiratory failure and corpulmonale, which would cause great damage to patients' health and quality of life. It has emerged as the fifth-largest burden in the global economy[1, 2], especially in developing countries, where COPD harbours high morbidity and mortality. One study has shown that about 1 million people die of COPD in China every year, accounting for 30% of the deaths of COPD across the world[3]. Also, COPD is the third and fourth leading cause of death in China's rural and urban areas, respectively[4]. Obviously, COPD has

become an important public health problem. It is imperative to comprehensively analyze its related factors and the complex relationship between them, and to take effective measures to prevent and reduce the occurrence of COPD as soon as possible.

Previous studies on COPD risk factors generally explored the risk factors of COPD by logistic regression, which required variables to be inter-independent, and it reflected the correlation between independent variables and outcome variables based on odds ratio. For example, in 2018, Wang Chen [5] studied the related risk factors of COPD using logistic regression, and found that gender, age, years of smoking and severe exposure to PM2.5 were significantly associated with COPD. However, in practical research, there is often a certain correlation between these influencing factors of disease. Therefore it fails to meet the prerequisite of independent variables of logistic regression. In addition, logistic regression is unable to reveal direct or indirect factors of COPD [6].

Bayesian networks(BNs) was firstly proposed by Pearl Judea in 1987, and then had been widely used[7]. Without strict statistical hypothesis[8], BNs construct a directed acyclic graph (DAG) to show the potential relationship among influencing factors, and use conditional probability distribution table (CPT) to reflect the correlation intensity among variables[9]. As such, BNs can directly show the complex network relationship between disease and influencing factors, and overcome the limitation of traditional logistic regression[10]. In addition, BNs can infer the probability of unknown nodes by using the information of known nodes, and flexibly show how the relevant risk factors make an impact on the risk of COPD[11, 12].

Bayesian networks learning refers to obtaining a complete Bayesian network by analyzing the existing information. The construction method consists of two types, parameter learning and structure learning[13]. Parameter learning assumes that the network structure is known, then determining the parameters in the network. This study focused on structure learning, which is more commonly used. Generally speaking, Bayesian networks structure algorithms can be divided into two parts, score-based search[14] and constraint-based algorithm[15].

The essence of score-based search is to find out the Bayesian networks structure whose score function reaches a maximum. Nonetheless, it's hard[16] to obtain an optimal network structure when the structure space becomes very large. The constraint-based algorithm boasts a high learning efficiency and can obtain the global optimal solution, but it also has some shortcomings. Firstly, independence of different nodes is sophisticated, and the number of independence tests between nodes increases exponentially with the increase of the number of nodes. Secondly, the results of the high-order conditional independence test are unreliable. Considering the limitations of the two kinds of algorithms, some scholars have proposed a hybrid algorithm. Max-Min Hill-Climbing(MMHC), adopted in this study, is a widely used hybrid algorithm, which includes two phases. In the first stage, it builds an undirected Bayesian network framework to reduce the search space by using constraint-based algorithms. Then the score-based search is used to add, delete and change the direction of edges in the constrained space to find the network with the highest score[16]. Thus, MMHC skillfully combines the two algorithms and effectively overcomes their shortcomings[17].

Undoubtedly, there are many related factors affecting COPD, if all of these factors are incorporated into the BNs, the network will become intricate, and some factors with weak correlation will reduce the accuracy of the model. Since Lasso regression does not take into account the correlation between features, it is not suitable for multiple collinear variables. Ridge regression cannot select the model, on account of no prediction factor with zero actual coefficients. However, Elastic Net[18] can combine the two and carries on the feature selection through cross-verification. By adopting Elastic Net, an ideal sparse model can be obtained and the influence of the correlation between the observed variables can be compensated.

Hence, we intended to employ Elastic Net to screen influencing factors from the original data, selecting factors with a strong correlation with COPD, and then used MMHC algorithm to build a network of COPD and related factors, exploring the potential relationship between COPD and these influencing factors, so as to provide a theoretical basis for clinical prevention and reduction of the occurrence of COPD.

Materials And Methods

Study participants

In this study, data were obtained from the COPD monitoring of residents from 2014 to 2015, which was carried out in Shanxi Province, China. After excluding missing data, 2072 valid cases were retained. Based on multi-stage stratified random sampling, a face-to-face survey was conducted among Chinese residents ≥ 40 years old in Taiyuan, Datong, Linfen and Xinzhou of Shanxi Province. Before the investigation, we obtained the support of the local neighborhood committee and the cooperation of study participants. The survey included basic information (such as gender, age, cultural level), respiratory symptoms (such as cough, expectoration, air hunger or dyspnea), personal diseases (such as childhood respiratory, hypertension) and risk factors exposure (such as household air pollution, occupational exposure). These factors and their assignments were depicted in Table 1.

The eligibility criterion for this study was residents of Chinese nationality aged 40 years or older who had lived in the monitoring area for at least 6 months in the 12 months prior to the survey. The exclusion criteria were shown below: (1) residents living in functional areas, such as barracks, military, student dormitories, nursing homes; (2) mental disorders or cognitive disorders (including dementia, comprehension impairment, deaf-mute); (3) tumor patients found and being treated; (4) high paraplegia; (5) pregnant or lactating women.

Table 1
(1) Factors and their assignments

Factors	Assignments	Population	Percent(%)
COPD(Y)	YES=1	277	13.4
	NO=2	1795	86.6
Gender(x_1)	Male=1	1073	51.8
	Female=2	999	48.2
Age(x_2)	≤49=1	759	36.6
	50~=2	726	35.0
	60~=3	484	23.4
	70~=3	103	5.0
Cultural level(x_3)	Junior high and below=1	1557	75.1
	Senior high school=2	391	18.9
	College diploma or above=3	124	6.0
BMI(x_4)	<18.5=1	16	0.8
	18.5~=2	675	32.6
	24.0~=3	909	43.9
	28.0~=4	472	22.8
COPD awareness (x_5)	YES=1	110	5.3
	NO=2	1962	94.7
Cough (x_6)	YES=1	183	8.8
	NO=2	1889	91.2

Table 1
(2) Factors and their assignments

Factors	Assignments	Population	Percent(%)
Expectoration (x_7)	YES=1	269	13.0
	NO=2	1803	87.0
Air hunger or dyspnea(x_8)	YES=1	381	18.4
	NO=2	1691	81.6
Childhood respiratory infections (x_9)	YES=1	37	1.8
	NO=2	2035	98.2
Respiratory disease (x_{10})	YES=1	263	12.7
	NO=2	1809	87.3
Cardiovascular and cerebrovascular diseases (x_{11})	YES=1	141	6.8
	NO=2	1931	93.2
Hypertension (x_{12})	YES=1	449	21.7
	NO=2	1623	78.3
Family history (x_{13})	YES=1	536	25.9
	NO=2	1536	74.1
Smoking status(x_{14})	YES=1	858	41.4
	NO=2	1214	58.6
Household air pollution(x_{15})	YES=1	1432	69.1
	NO=2	640	30.9
Occupational exposure (x_{16})	YES=1	903	43.6
	NO=2	1169	56.4

Quality control

To ensure the reliability and validity of data, strict measures had been taken in this study. The investigators received standardized professional training before the survey. After passing the investigation, they conducted a face-to-face survey on participants and used questionnaires to collect relevant data. On-site and remote quality control were implemented through synchronous recording. All measuring instruments were calibrated before measurement. All data were entered twice into a database and checked for errors or omissions.

Elastic Net

Regularization is a technique for adding penalties to the objective function. This penalty controls the complexity of the model by reducing the value of the regression coefficient. Elastic Net[18]is a linear regression model with L1 and L2 norms as regularization matrix. Not only does it retain the characteristic to easily produce feature sparsity like Lasso method, but also inherits the stability of ridge regression. Its algorithm formula is as follows.

$$\hat{\mathrm{B}} = {}_{\beta}^{\mathrm{argmin}}\frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|_2^2 + \lambda(\alpha\|B\|_1 + \frac{(\vdash a)}{2}\|B\|_2)$$

As we can see, λ represents the penalty coefficient and β is the regression coefficient. For the convex combination of L1 and L2 regularization, the l1_ratio parameter is used for adjustment. The final value of the parameter is selected by ten-fold cross-validation to select the parameter value with the lowest model error.

Bayesian networks

Bayesian networks is a probability graph model, which can show the probability dependence intensity between factors. It is a directed acyclic graph based on probability theory and graph theory, which consists of nodes representing the variables U = {xi, …, $\mathrm{x_n}$} and the directed edges represent the relationship between variables [12]. If the edge from $\mathrm{x_i}$ to $\mathrm{x_j}$ exists, then $\mathrm{x_i}$ is the parent of $\mathrm{x_j}$and $\mathrm{x_j}$ is the child of $\mathrm{x_i}$ [13]. Each node can quantitatively describe the probability correlation between the node and its parent node through the attached conditional probability distribution table (CPT). In BNs, the formula for calculating the joint probability distribution function of all nodes is as follows.

$$\mathrm{P}\left(\mathrm{x}_1 \square \mathrm{x}_2 \square \cdots \square \mathrm{x}_\mathrm{n}\right) = \mathrm{P}\left(\mathrm{x}_1\right)\mathrm{P}\left(\mathrm{x}_2 \middle| \mathrm{x}_1\right)\cdots\mathrm{P}\left(x_{n|x_1,x_{2,}\cdots,x_{n-1}}\right)$$

$$= \Pi_1^{\mathrm{n}}\mathrm{P}\left(\mathrm{x_i}\middle|\pi\left(\mathrm{x_i}\right)\right)$$

$\pi\left(\mathrm{x_i}\right)$ is the set of parent nodes of $\mathrm{x_i}$,$\pi\left(\mathrm{x_i}\right) \subseteq \square \mathrm{x}_1, \ldots, \mathrm{X}_{\mathrm{i}-1} \square$.

MMHC

MMHC algorithm is a widely used Bayesian network hybrid structure learning algorithm, which is mainly divided into two stages. In the first stage, the MMPC algorithm is employed. It can determine the existence of edges without direction, from which the Bayesian network can be constructed. The MMPC algorithm also includes two phases, the first phase starts from the empty set, and then variables input CPC into the empty set successively by using the max-min heuristic function. The first phase doesn't end until all remaining nodes are independent of target node; in the second phase, false positive nodes were

deleted through the conditional independence test. For a subset of variables S (S⊆CPC), if Ind (X, T | S) was true, X was deleted from CPC.

In the second stage of the MMHC algorithm, the mountain climbing method is used to locally adjust the current model by adding, deleting and changing the direction of the edges, so as to get several undetermined models, and then calculate the score of each undetermined model to obtain the Bayesian network with the highest score [15].

Definition

The ratio of forced expiratory volume in the first second (FEV1) to forced vital capacity (FVC) < 70% after the bronchodilation test was determined as COPD patients. The age consisted of four types: 40-49, 50-59, 60-69, ≥ 70, and cultural level was divided into three levels: junior high school and below, senior high school, college and above. Bodyweight was classified as: underweight (BMI < 18.5kg/m2); Normal body weight (BMI 18.5-23.9kg/m2); Overweight (BMI 24.0-27.9kg/m2); Obesity: (BMI≥ 28.0kg /m2). Participants who smoked more than one cigarette a day for the past six months were defined as smokers. The use of wood, animal manure or coal for cooking or heating over the past six months or more had been defined as household air pollution. Exposure to dust or harmful gases at work (including farm work) was defined as occupational exposure. One or both parents who had suffered from respiratory diseases such as asthma, chronic bronchitis, emphysema, were defined as a family history of respiratory diseases.

Statistical analysis

Statistical description and analysis of influencing factors of COPD were analyzed by IBM SPSS Version 22. Elastic Net feature screening was carried out with the ElasticNetCV program in Sklearn linear_model library in Python 3.7.0 and CV was set to 10. The structure of BNs was constructed by the MMHC function of bnlearn package in R studio 4.0.5 software, and the maximum likelihood estimation method was used for parameter learning. The drawing of the BNs graph and CPT were realized by Netica software.

Results

Characteristics of the study population

Among the 2424 initial study participants, 352 subjects with incomplete data were excluded, and 2072 participants were finally taken into analysis. Among them, 51.8% were men and 48.2% were women. 36.6% of the participants were between 40 and 49, 35.0% were between 50 and 59, 23.4% were between 60 and 69, and 5.0% were over 70 years old. (As shown in Table 1) In 2014, the prevalence of COPD among residents 40 years and older in Shanxi Province, China represented 13.4% (male 19.9%, female 6.3%). With the increase of age, the prevalence of COPD also increases gradually. The highest prevalence of COPD among people older than 70 years occupied 22.3%, as shown in Figure 1.

COPD related factors screening using Elastic Net

16 risk factors related to COPD were included in the Elastic Net model, and the key parameter values (λ = 0.18595, α = 0.12) optimizing the model performance were selected by a ten-fold cross-validation method. In the end, the coefficients of influencing factors not closely related to COPD would be compressed to 0 and eliminated, and the final 10 variables (Table 2) were obtained. This method was used to determine the strong correlation factors affecting COPD, simplifying the structure of the later BNs.

Table 2
Selected variables and regression coefficients

Variable	coefficients	Variable	coefficients
x_1	0.24111383	x_8	0.07155523
x_2	-0.07398471	x_{10}	0.16883457
x_4	0.074471365	x_{13}	0.04980561
x_6	0.22663873	x_{14}	0.22661544
x_7	0.13507333	x_{15}	0.08082141

Bayesian networks model of COPD

According to the 10 factors related to COPD selected by Elastic Net in the previous stage, the MMHC algorithm was used to further construct the BNs model of COPD and its related factors. As shown in Figure 2, a COPD model with 11 nodes and 18 directed edges was constructed. The directed edges represented the dependence of various related factors on COPD. The numbers in the figure represented the prior probability of each node. For example, the prior probability of COPD was 0.134, that is, P(COPD)=0.134. The results showed that smoking status, household air pollution, family history, cough, air hunger or dyspnea were directly related to COPD. Among them, smoking status, household air pollution, and family history constituted the parent nodes of COPD, that is, they were related to the occurrence of COPD. Cough, air hunger or dyspnea were child nodes of COPD. Namely, COPD was related to the occurrence of Cough, Air hunger or dyspnea.

Reasoning model of COPD

BNs can infer the probability of an unknown node (COPD) based on the state of a known node, and make COPD risk determination possible. If an individual smokes, the probability of suffering from COPD is 0.215, that is, P(COPD | Smoking status) = 0.215, as shown in Figure 3; if the individual has used wood, animal feces or coal in the past 6 months or more Cooking or heating, the probability of suffering from COPD becomes 0.246, that is, P(COPD | Smoking status, Household air pollution) = 0.246, as shown in Figure 4; if the individual has a family history of respiratory disease at the same time, then the possibility of suffering from COPD is 0.280, that is, P(COPD | Smoking status, Household air pollution, Family history)=0.280, as shown in Figure 5. When a body suffers from COPD, its usual probability of Cough

rises from 0.0887 to 0.201, that is, P (Cough | COPD) = 0.201, and its probability of Air hunger or dyspnea rises from 0.184 to 0.289, that is, P (Air hunger or dyspnea | COPD) = 0.289, as shown in Figure 6.

Table 3
COPD conditional probability table

Family history	Smoking status	Household air pollution	COPD	
			NO	YES
Yes	No	No	93.939	6.061
Yes	No	Yes	80.928	19.072
Yes	Yes	No	82.609	17.391
Yes	Yes	Yes	71.978	28.022
No	No	No	95.775	4.225
No	No	Yes	93.488	6.512
No	Yes	No	86.225	13.775
No	Yes	Yes	76.611	23.389

Discussion And Conclusions

Amid ageing population in China, COPD has become an important public health issue. Globally, it is the main cause of disability among elderly population and has become the fifth-largest burden of the global economy [2]. This study showed that the prevalence of COPD in Shanxi Province, China in 2014 was 13.4%, which was similar to the national COPD prevalence of 13.6%. However, in the past ten years, the prevalence of COPD among residents over 40 in China has increased from 8.2% in 2002[19] to 13.7% in 2012[20]. This showed that Shanxi Province should attach importance to the prevention and treatment of COPD.

The BNs constructed by the MMHC algorithm can explore the complex network connections between COPD and its various influencing factors. The results of BNs model showed that P(COPD)=0.134. Smoking status, household air pollution, and family history were directly related to COPD, and gender was indirectly related to COPD through smoking. In addition, the BNs can also describe the relationship between other factors, such as the network relationship between family history, respiratory disease, air hunger or dyspnea, cough, expectoration and other factors, as shown in Figure 2. Logistic regression can't show relationships between variables, because it is a model built on the condition that these factors are inter-independent. Table 3 is the conditional probability distribution table of the parent node of COPD. It can be seen that the probability dependence between COPD and the three-parent nodes of smoking status, household air pollution, and family history. If an individual had smoking status, household air

pollution, family history at the same time, then he was 28.0% likely to develop COPD, with P (COPD | smoking status, household air pollution, family history) =0.280.

Smoking is currently recognized as the most important risk factor for COPD. The chemicals and fine particles produced during tobacco burning are the main causes of chronic bronchial inflammation and airway obstruction. Su J et al.[21] found that in the Joint association of cigarette smoking and PM with COPD among urban and rural adults in regional China, after adjusting for other factors, the risk of COPD for smokers is 2.46 times more than that of non-smokers. In 2014, the smoking rate of residents aged 40 and over in Shanxi Province reached 41.4%. Male smokers exceeded 70%, reflecting the high prevalence of smoking behavior among the population in our province. In terms of COPD prevention and control, tobacco control and non-exposure to tobacco smoke prove one of the most important intervention methods.

Pollutant fuels refer to biofuels (wood, animal manure, charcoal, firewood, crop waste), coal and kerosene fuels, etc. Household air pollution refers to households using biofuels, coal fuels and other polluting fuels for cooking and heating. In 2016, WHO[22] estimated about 3.1 billion people in low-and middle-income countries still use contaminated fuel for cooking, causing about 4.3 million premature deaths each year, equivalent to 7.7% of global deaths, and causing one-third of low-and middle-income countries death from COPD. In 2014, the WHO[23] issued the "Guidelines for Indoor Air Quality-Executive Summary of Household Fuel Combustion", strongly recommending that untreated coal should not be used as household fuel, and households are not encouraged to use kerosene. In 2014, households of residents aged 40 and over in Shanxi Province used polluted fuels for cooking and heating. The household air pollution rate reached 69.1%. Therefore, sufficient attention should be paid to the indoor pollution caused by the use of biofuels and other polluting fuels in households to prevent and control COPD.

Having a family history of respiratory diseases will increase the incidence of COPD, suggesting that genetic susceptibility is also closely related to the incidence of COPD. At present, some studies have found that the polymorphisms of α-antitrypsin, matrix metalloprotein, tumor necrosis factor α, interleukin and other genes were related to the pathogenesis of COPD, but further research is needed to clarify[24–26].

To sum up, building BNs based on Elastic Net can further figure out complex network relationships with linkage effects between factors based on finding strong disease-related factors, which is more intuitive to reveal the network connection between disease and related factors. After fully understanding the network connections between diseases and factors, more targeted measures should be taken into disease prevention and control.

Abbreviations

COPD: Chronic Obstructive Pulmonary Disease, MMHC: Max-Min Hill-Climbing, DAG: Directed Acyclic Graph, CPT: Conditional probability distribution table, BNs: Bayesian Networks

Declarations

Funding Information

This study was supported by the national natural science foundation of China project (grant numbers 81973155). The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Acknowledgements

We thank all teachers in the statistical research office of Shanxi medical university. Authors would also like to acknowledge all interviewers for survey data collection work.

Ethics approval and consent to participate

The study was organized by the National Health Commission Disease Control and Prevention Bureau. Informed consent was signed by all study participants or their agents. All methods were carried out in accordance with relevant guidelines and regulations, and all experimental protocols were approved by a named institutional and/or licensing committee.

Consent for publication

Not applicable

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests:

The authors declare no competing interests.

Author's contributions

DCQ, JHR and HR analyzed and interpreted the data, and are major contributors in writing the manuscript. LQLH were responsible for preprocessing the data and checking the results. ZPR conducted the survey and collected data. XCW, MCL and YCQ were involved in compiling the data and summarizing the results. LXQ gave constructive suggestions for the manuscript. All authors read and approved the final manuscript.

References

1. MARTINEZ F J, HAN M, ALLINSON J P, et al. At the Root: Defining and Halting Progression of Early Chronic Obstructive Pulmonary Disease [J]. 2018, 196(12): 569.

2. SONG Q, CHEN P, LIU X M. The role of cigarette smoke-induced pulmonary vascular endothelial cell apoptosis in COPD [J]. Respiratory Research, 2021, 22(1).
3. YIN P, WANG H, VOS T, et al. A subnational analysis for mortality and prevalence of chronic obstructive pulmonary disease in China 1990- 2013: Findings from Global Burden of Disease Study (GBD) 2013 [J]. 2016, 1269-80.
4. ZHU B, WANG Y, MING J, et al. Disease burden of COPD in China: a systematic review [J]. 2018, 13(1353-64.
5. WANG C, XU J, YANG L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study [J].
6. ALLISON P D J S P. Logistic regression using the SAS system: Theory and application [J]. 1999,
7. KOCH D, EISINGER R S, GEBHARTER A J J O T B. A causal Bayesian network model of disease progression mechanisms in chronic myeloid leukemia [J]. 2017, 433(94-105.
8. NI W Q, LIU X L, ZHUO Z P, et al. Serum lipids and associated factors of dyslipidemia in the adult population in Shenzhen [J]. Lipids in health and disease, 2015, 14(71.
9. WEI Z, ZHANG X L, RAO H X, et al. [Using the Tabu-search-algorithm-based Bayesian network to analyze the risk factors of coronary heart diseases] [J]. Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi, 2016, 37(6): 895-9.
10. KUNG-JENG, WANG, BUNJIRA, et al. Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: A case study of Taiwan [J]. 2014,
11. BURNSIDE E S, RUBIN D L, SHACHTER R D. Using a Bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography [J]. Studies in health technology and informatics, 2004, 107(Pt 1): 13-7.
12. HUGHES R E. Using a Bayesian Network to Predict L5/S1 Spinal Compression Force from Posture, Hand Load, Anthropometry, and Disc Injury Status [J]. Applied bionics and biomechanics, 2017, 2017(2014961.
13. KAEWPRAG P, NEWTON C, VERMILLION B, et al. Predictive models for pressure ulcers from intensive care unit electronic health records using Bayesian networks [J]. 2017, 17(S2): 65.
14. CAMPOS L. Independency Relationships in Singly Connected Networks [J]. 1994,
15. HECKERMAN D, GEIGER D, CHICKERING D M J M L. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data [J]. 1995, 20(3): 197-243.
16. TSAMARDINOS I, BROWN L E, ALIFERIS C F J M L. The max-min hill [J]. 2006, 65(1): 31-78.
17. HAFF I H, AAS K, FRIGESSI A, et al. Structure learning in Bayesian Networks using regular vines [J] Computational Statistics and Data Analysis . 2016, 186-208
18. ZOU H, HASTIE T. Addendum: "Regularization and variable selection via the elastic net" [J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005), no. 2, 301–320, MR2137327] [J]. journal of the royal statistical society, 2010, 67(5): 768-.

19. ZHONG N, CHEN W, YAO W, et al. Prevalence of Chronic Obstructive Pulmonary Disease in China: A Large, Population-based Survey [J]. American Journal of Respiratory and Critical Care Medicine, 2007, 176(8): 753-60.
20. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study [J]. Lancet, 2018,
21. SU J, YE Q, ZHANG D, et al. Joint association of cigarette smoking and PM2.5 with COPD among urban and rural adults in regional China [J]. BMC Pulmonary Medicine, 2021, 21(1):
22. ORGANIZATION W H. Burning opportunity: clean household energy for health, sustainable development, and wellbeing of women and children [J]. 2016,
23. ORGANIZATION W H. WHO Guidelines for indoor air quality: household fuel combustion [J]. Who Guidelines for Indoor Air Quality Dampness & Mould, 2014,
24. CLANCY J, TURNER C. Smoking and COPD: the impact of nature-nurture interactions [J]. British Journal of Nursing, 2013, 22(14): 820, 2.
25. CUI K, GE X Y, MA H L. Association of the TNF-α +489 G/A polymorphism with chronic obstructive pulmonary disease risk in Asians: Meta-analysis [J]. Genetics and molecular research: GMR, 2015, 14(2): 5210-20.
26. SAPEY E, WOOD A M, AHMAD A, et al. Tumor necrosis factor-{alpha} rs361525 polymorphism is associated with increased local production and downstream inflammation in chronic obstructive pulmonary disease [J]. American Journal of Respiratory & Critical Care Medicine, 2010, 182(2): 192-9.

Figures

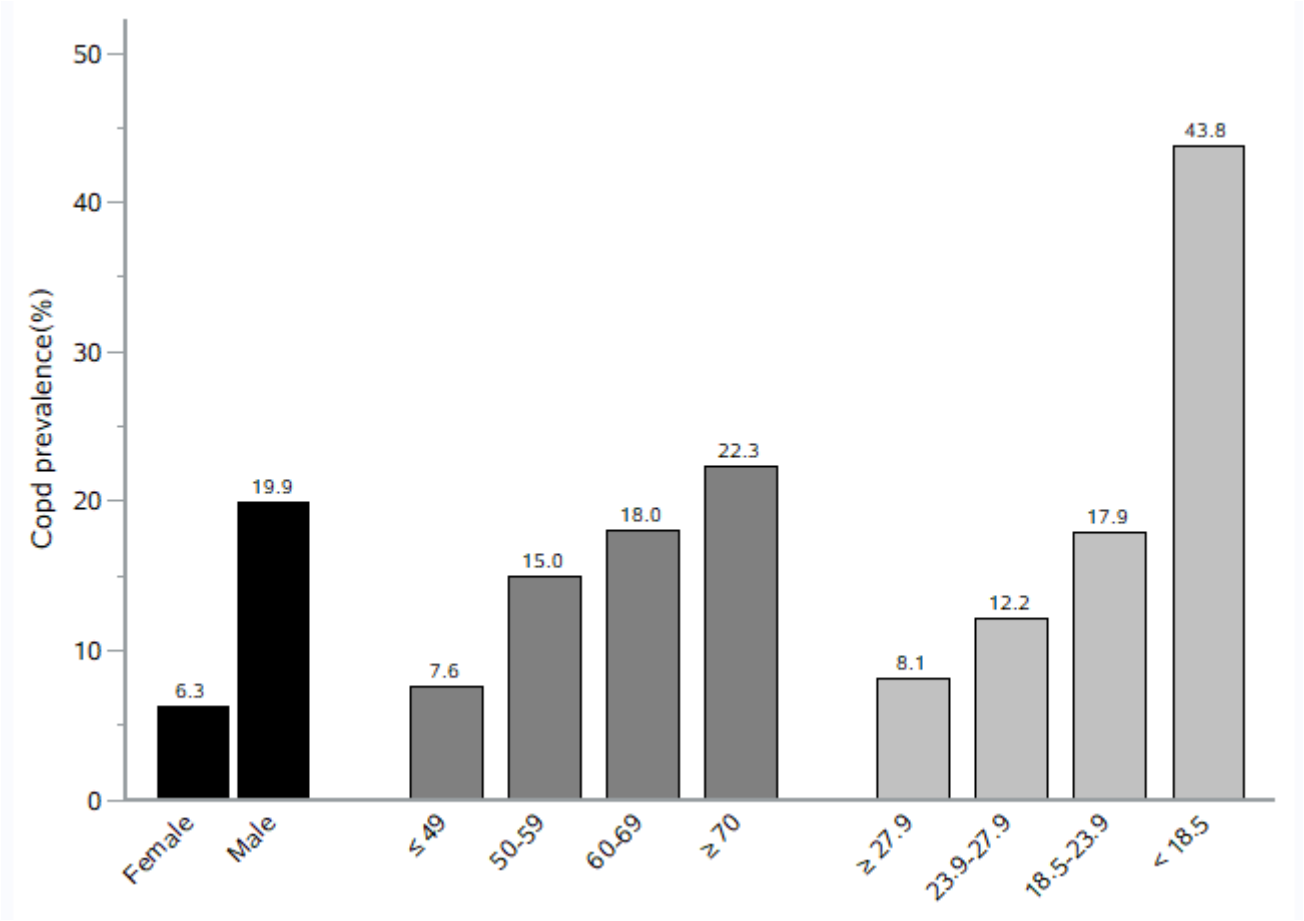

Figure 1

The prevalence of COPD in different gender, age and BMI

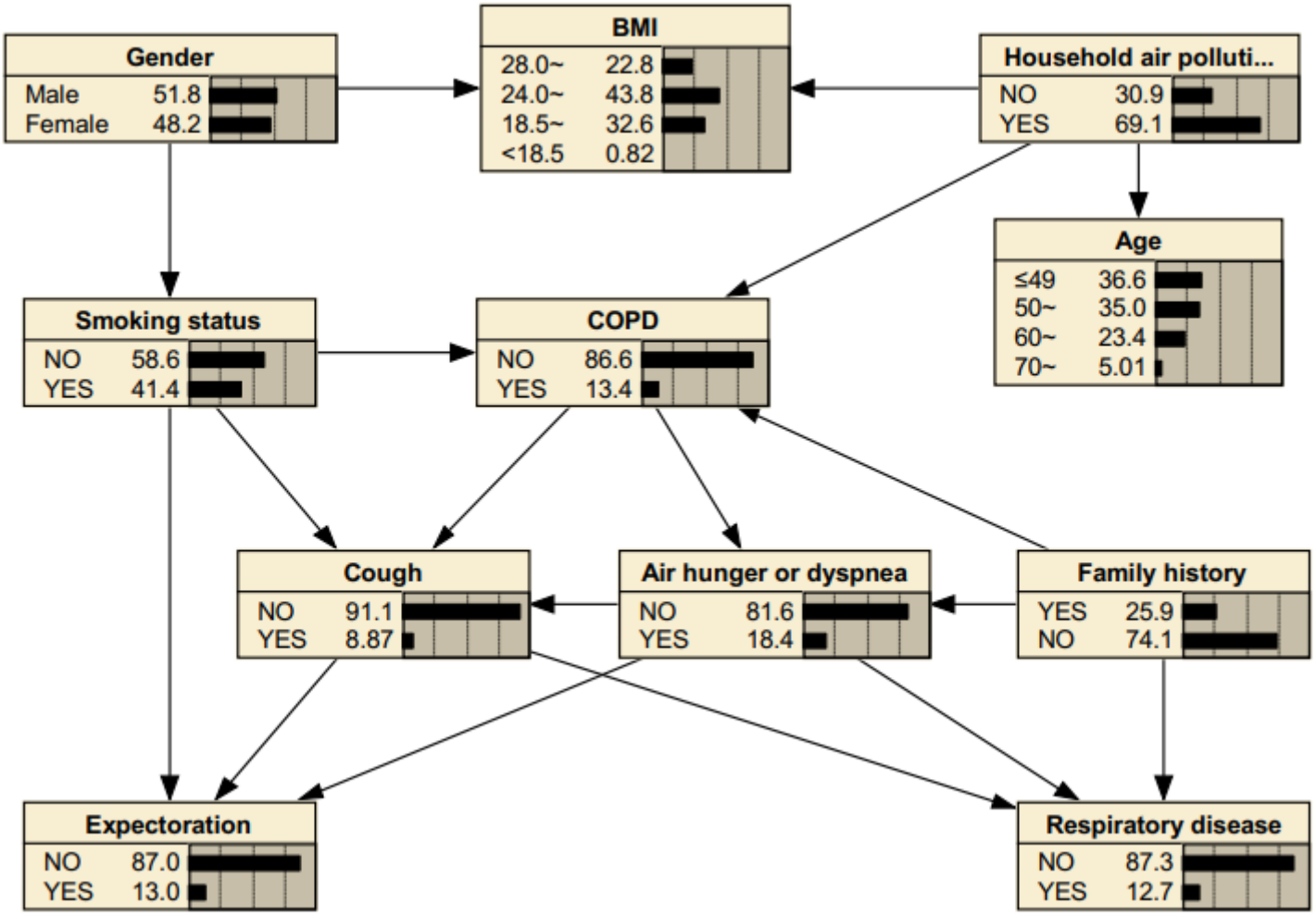

Figure 2

MMHC algorithm to construct COPD Bayesian networks and prior probability

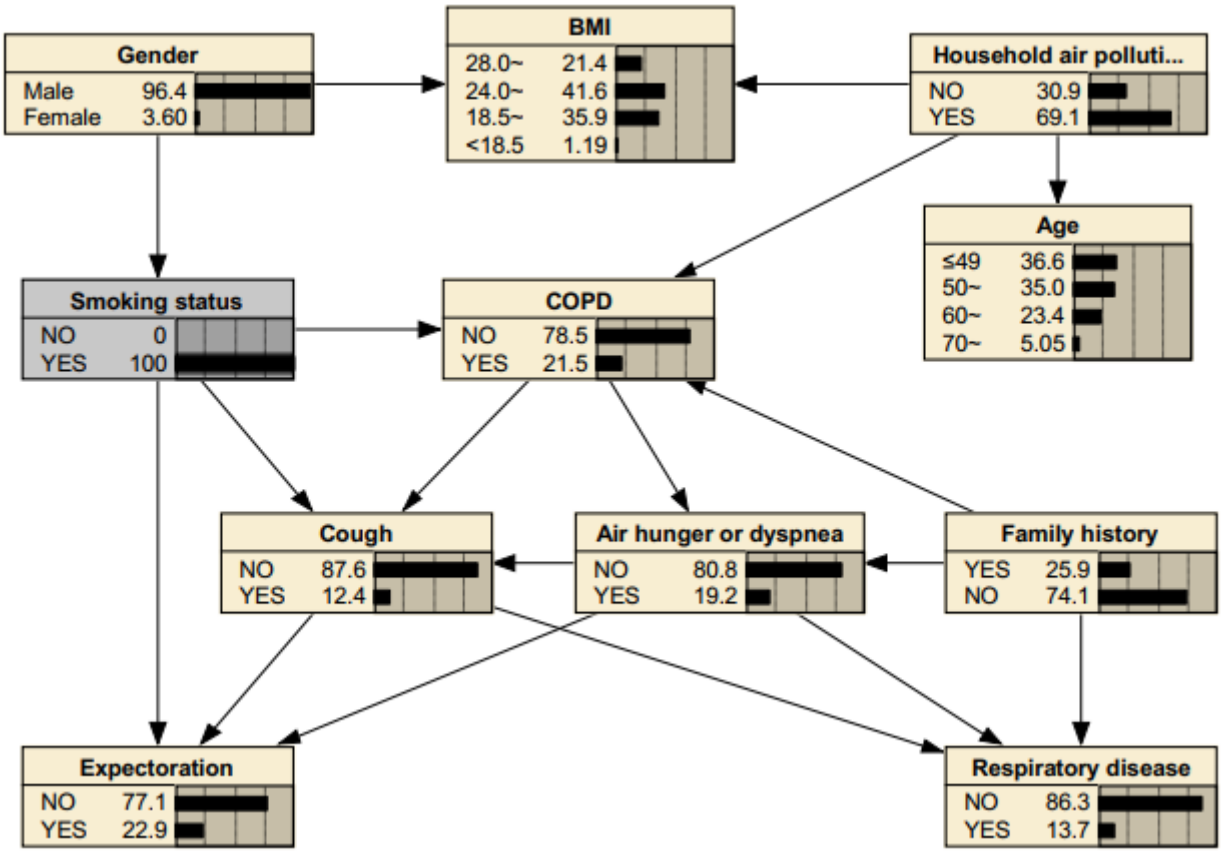

Figure 3

The Bayesian networks under known evidence variables. The figure was plotted using Netica (www.norsys.com).

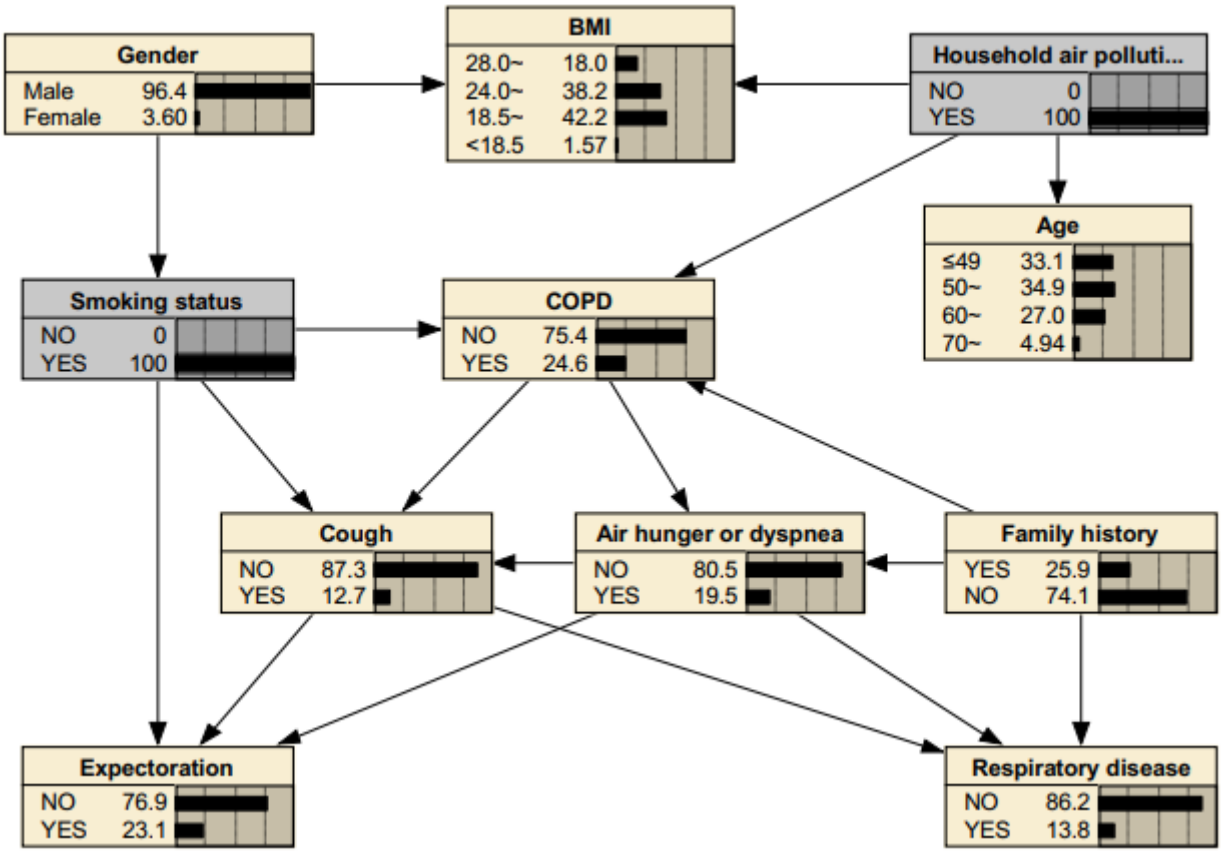

Figure 4

The Bayesian networks under known evidence variables. The figure was plotted using Netica (www.norsys.com).

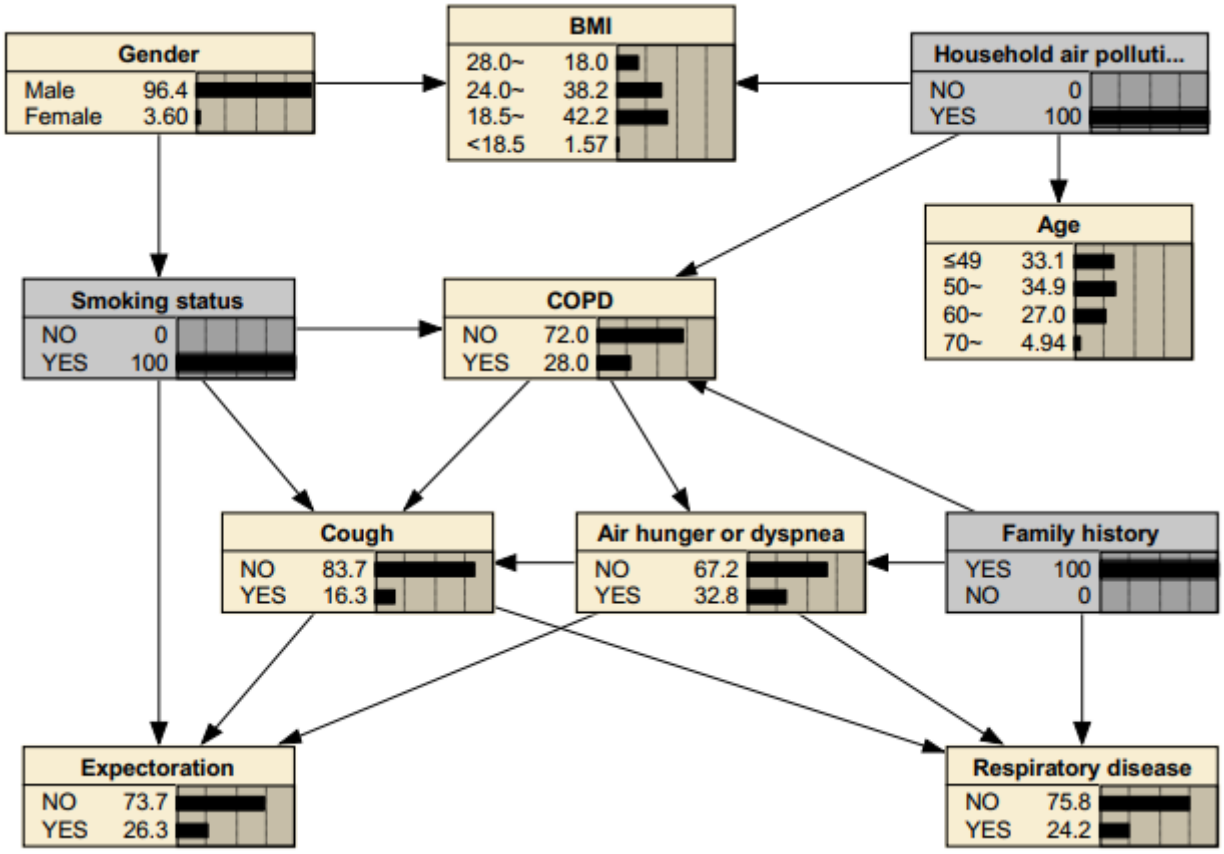

Figure 5

The Bayesian networks under known evidence variables. The figure was plotted using Netica (www.norsys.com).

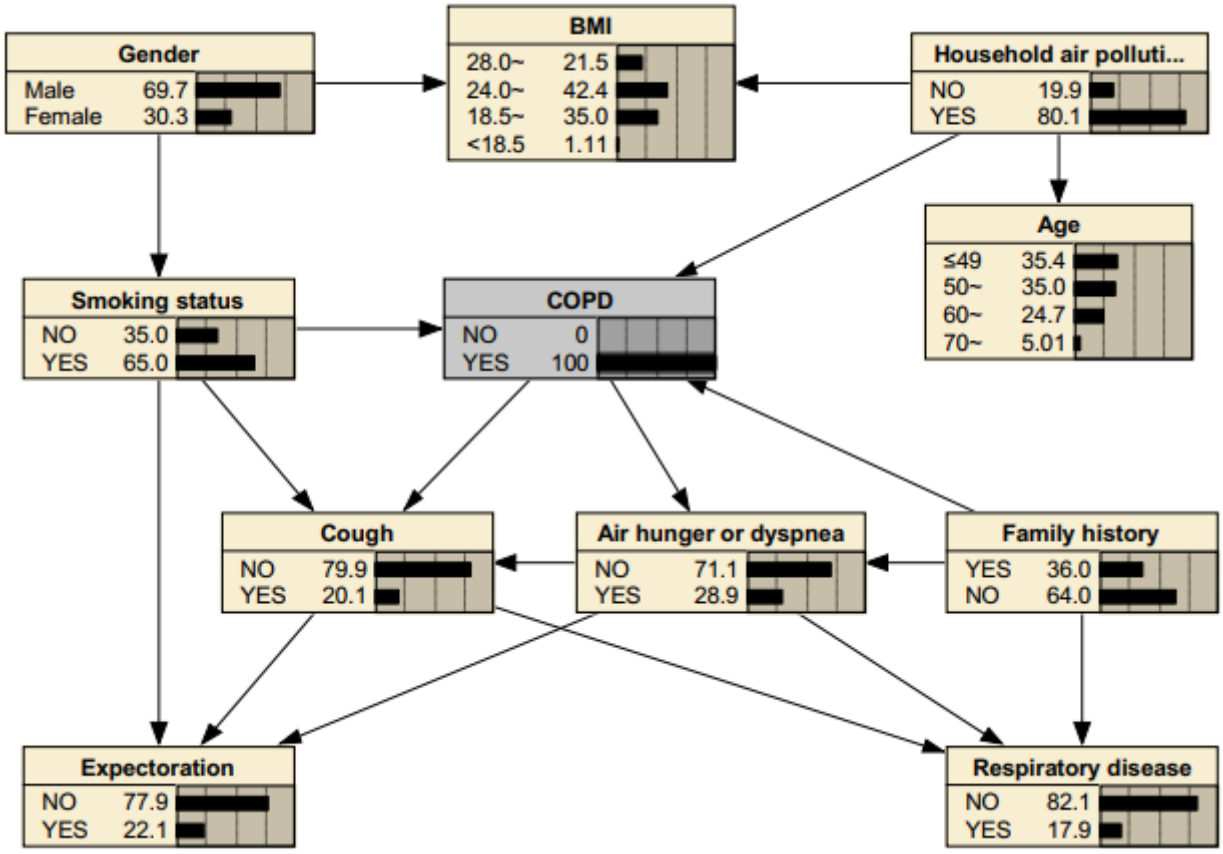

Figure 6

The Bayesian networks under known evidence variables. The figure was plotted using Netica (www.norsys.com).