Original Article

A data mining algorithm for determination of influential factors on the hospitalization of patients subject to chronic obstructive pulmonary disease

Mohammad Mozafarinia[1], Ali Reza Shahriyari[1*], Mohamad Karim Bahadori[1], Ali Ghazvini[1], Seyyed Shamsadin Athari[2*]

[1]*Baqiyatallah University of Medical Sciences, Tehran, Iran*
[2]*Department of Immunology, School of Medicine, Zanjan University of Medical Sciences, Zanjan, Iran*

Received: 2 April, 2018; Accepted: 19 July, 2018

Abstract

Background: The present study is on the development of a data mining algorithm for finding the influential factors on the hospitalization of patients subject to chronic obstructive pulmonary disease. **Materials and Methods:** This is a descriptive analytical study conducted cross sectionally in 2017 on a research community of 150 people with disease symptoms referred to clinics and hospitals across Tehran (Iran). The people were surveyed by a self-designed questionnaire, including queries on life style and family information. The sampling was simple intuitive from previously published studies. The modeling of the data was based on the CRISP method. The C5 decision tree algorithm was used and the data was analyzed by RapidMiner software. **Results**: The common symptoms of the patients were found to be shortness of breath, cough, chest pain, sputum, continuous cold, and cyanogens. Besides, the family history, smoking, and exposure to allergic agents were other influential factors on the disease. After accomplishment of this study, the results were consulted with the experts of the field. **Conclution:** It is concluded that data mining can be applied for excavation of knowledge from the gathered data and for determination of the effective factors on patient conditions. Accordingly, this model can successfully predict the disease status of any patient from its symptoms.

Keywords*:* Chronic lung obstruction; data mining; decision tree algorithm; disease status of patient; RapidMiner

***Corresponding Author:** Ali Reza Shahriyari; Seyyed Shamsadin Athari, Email: SS.Athari@gmail.com.

Please cite this article as: Mozafarinia M, Shahriyari A R, Bahadori M K, Ghazvini A, Rahmani M, Athari S Sh. A data mining algorithm for determination of influential factors on the hospitalization of patients subject to chronic obstructive pulmonary disease. Arch Med Lab Sci. 2018;4(2): 30-35.

Introduction

The continuous rise in urban air pollution of industrialized cities led to several respiratory diseases in people of different age groups. Currently, about 329 million people are subject to this disease that is amount to 5% of the world population. It was the third cause of death in 2012, more than 3 million patients [1]. The diseases relevant to respiratory systems could disturb the correct function of the lung that characterized by long term breathing problems and poor airflow [2]. The occurrence of this problem is rising continuously and ignorance of safety issues could also exacerbate the problem. The chronic obstructive pulmonary disease (COPD) became one of the common health problems in different age groups. The COPD is a new term for such problems relevant to chronic obstruction of airways such as emphysema, chronic bronchitis, asthma, and any combination of these disorders. The chronic bronchitis and emphysema occur by inflammation on bronchus and over inflation of the alveoli.

Treatment of the COPD can inhibit symptoms and minimize any further problems; although the damage to lungs due to the COPD is permanent. There is no absolute cure for this disease, but the correct control of it can only be done by its correct diagnosis at its early status. Nowadays, new techniques such as data mining come to the spotlight of medical research. This technique was introduced as one of top ten technologies in current era [3]. Data mining is the process of generating valid unknown, understandable, and trustworthy knowledge from data. It can find novel, accurate, and understandable patterns between data so as to generate knowledge behind massive data. Actually, the data mining lighten the patterns that were indiscernible for researchers by their own. In data mining, the decision tree was introduced as a common method to resolve many questions. For example, Cano et al. [4], were applied data mining techniques to generate a COPD knowledge base from clinical and experimental data. This knowledge resource allows users to retrieve specific molecular networks from general clinical and experimental data sources. Cristobalina et al. [5], conducted a cross-sectional epidemiological study on 402 individuals using five different data mining techniques. They find out that dyspnoea is the most influential factor, allowing early diagnosis of people with COPD. Sanchez-Morillo et al. [6], also conducted a review of different data mining algorithms and factors associated with their performance in early detection of COPD and asthma. In this systematic review, they insist on the need for further research to develop new algorithms with improved predictive capability and clinical reliability.

In the present research, the data of COPD from people with disease symptoms referred to clinics and hospitals across Tehran (Iranwere analyzed by C5 decision tree technique. This analysis was conducted in RapidMiner environment in order to reach a new algorithm for determination of influential factors on the hospitalization of patients subject to COPD. This new predictive model can lead us to a pattern for prediction of interventions in the COPD.

Methods

The success of data mining technique in the development of a predictive model highly depends on the comprehensiveness of the selected methodology. In the present work, the CRISP method was applied as it contains all the steps from data understanding, data preparation, modeling, evaluation, and deployment, Fig. 1.

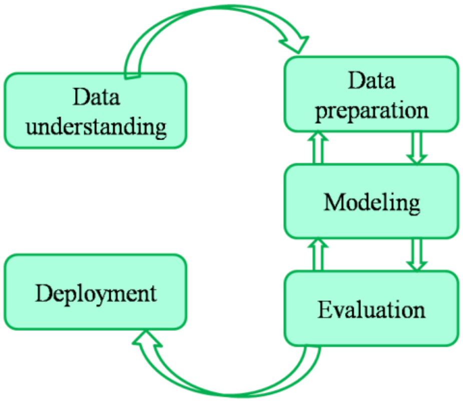

Figure 1. Process diagram showing the five phases of CRISP data mining.

Data understanding. In this step, understanding of the data and determination of data mining goals was considered in order to develop a successful model for accurate and rapid diagnosis of patient conditions.

Data collection and description. The data was collected from the records of hospitalized patients. To this end, the records of referred patients to specialized clinics on COPD across Tehran (Iran) were considered. In some cases, the patients were interviewed and questionnaires were completed. There were 200 filled questionnaires from both male and female patients during a 6 month period in 2017. The studied characteristics were 50 queries on personal and family information as well as their clinical information, Table 1. The characteristics were categorized as numerical and nominal ones and defined as binary inputs into the model.

Data preparation. The data preparation was conducted at first by removing the initial and surname of the person, record number in the hospital, and their contact information. The incomplete information of any individual was removed later in the modeling process by defining a block diagram in the algorithm.

Reliability of the questionnaire. The reliability is the extent to which a questionnaire

Table 1. The studied characteristics of the COPD patients and their respective category

Characteristic	Categorical Type	Characteristic	Categorical Type
Age	Numerical	Wheezing chest	Nominal
Gender	Nominal	Shortness of breath	Nominal
Job	Nominal	Restlessness and confusion	Nominal
Length	Numerical	Cyanogens	Nominal
Weight	Numerical	Exposure to allergic agents	Nominal
Strong attacks	Nominal	Other allergic agents	Nominal
Repeated attacks	Nominal	Family history	Nominal
Emergency need	Nominal	Inability to talk	Nominal
Hospitalization	Nominal	Aspirin usage	Nominal
Reception in priority	Nominal	Pregnant women	Nominal
Lung abscess	Nominal	Addicting agent consumption	Nominal
Cough with sputum	Nominal	Cough cold	Nominal
Bad breath	Nominal	Stress and anxiety	Nominal
Sweating	Nominal	Immunodeficiency disease	Nominal
Fever	Nominal	Work absent	Nominal
Weight loss	Nominal	Sleep disorder	Nominal
Chest pain	Nominal	Disease awareness	Nominal
Anesthesia	Nominal	Wheezing with activity	Nominal
Smoking	Nominal	Respiratory volume reduction	Nominal
Antibiotic use	Nominal	Fatigue	Nominal
Cough	Nominal	Choking sense	Nominal
Exposure to smoke	Nominal	Chest tightness	Nominal
Phlegm production	Nominal		

produces similar results on repeated trials. The Cronbach's alpha was used for the reliability measurement. The values between 0.7 and 0.9 show a good measure of consistency. The measured value for the present work was 0.8 that is good. Accordingly, the questionnaire is reliable and can be applied for further analysis.

Validity of the questionnaire. Validity shows how much the measurements are close to the real value. It can be evaluated by different means and in the present work it is based on the knowledge of the experts. To be sure of the questionnaire validity, it is revised after taking the comments of eleven expert physicians.

Modeling. At first step of modeling, the impact rate of each characteristic (Table 1) was considered by SMC programming in Matlab environment. The modeling was done in RapidMiner software. A predictive model should be used to determine the conditions of the patients subject to COPD. The decision tree algorithm was used for this model creation. As the goal set has two normal and bad conditions, the C5 decision tree algorithm was used because of working with binary sets. The inputs of the algorithm were 50 characteristics of the questionnaire and the output was the patient conditions.

The original data set is split into two training and test sets with the respective percentages of 30% and 70%. The models were generated by the 10 split method such that the data set is randomly divided into 10 parts [7]. Subsequently, each one of these 10 parts is chosen as the test data and the nine others as training data. This strategy was chosen mainly because it would result in the most accurate predictions. The class labels were chosen as shown in Table 2 and later by the use of SMC method, the main factors in occurrence and severances of the disease were chosen as the tree branches. The branching divisions were based on Gini index such that each tree is split into sub branches so as to have the data of each node in a class. The leaf to root path was then passed in the reverse direction and the generated rules are expressed conditionally.

Evaluation. The evaluation of the model results is an essential step in order to apply it in current form or improve it. To evaluate model accuracy, the data were divided into three training, test, and validation sets. The training set was for the design of decision tree and the test data was for tree evaluation and assigning a class label. The validation set were set to test the correctness of the model. There are various indexes for evaluation of the correctness of the results from classification methods, as allergy, transparency, accuracy, and precision, among others [8, 9]. The accuracy of each classification of training data set is the percentage of training set observations that is correctly classified by the method. The test data were used for calculation of this index. The precision of the model was evaluated by the perturbation matrix, Table 2. If data was set in M class, a class matrix has a least size of M×M. The ideal case is that one with most of the data relevant to observations on the main diagonal of the matrix and the rest of the matrix values be nearly or equal to zero.

Table 2. The perturbation matrix for evaluation of the model precision

Class	**Prediction of classification algorithm**	
Positive	True positive (TP)	False negative (FN)ccccc
Negative	False positive (FP)	True negative (TN)

TN stands for the number of records with negative class and the classification algorithm correctly assigned their class negative; TP for the records with positive class and the classification algorithm correctly assigned their class positive; FN for the records with positive class and the classification algorithm assigned their class negative; FP for the records with negative class and the classification algorithm assigned their class positive.

The precision or rate of classification algorithm is the best performance metrics that is measure total accuracy of a classification. This criterion is the commonly used metrics for evaluation of the performance of classification algorithms. It shows that the designed classification to what extent correctly predict the test records. The classification precision can be calculated by the following equation:

$$CA = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \quad (1)$$

Where, it shows that the TP and TN values are the prominent ones to be optimized in a two class problem. The results of the model accuracy evaluation were presented in Table 3. Using the data of asthma patients, this model can predict their conditions with 100% accuracy. From all hospitalized people, the 99 individuals were of COPD disease. The COPD diagnosis was correctly done for all patients except two cases. Accordingly, it could diagnosis the COPD with 100 % accuracy. The other 51 cases were hospitalized for other disease as all correctly diagnosed.

Table 3. Accuracy test results of the algorithm in diagnosing the COPD

	True Y	**True N**	**Class precision**
Predicted Y	99	0	100.00
Predicted N	0	51	100.00
Class recall	100.00	100.00	

Results

All the effective factors on COPD including personal information, family history, life style, and clinical information of the disease were considered in the present work. The impact rate of the most important symptoms of the COPD was presented in Table 4. The most important symptoms of the disease were determined shortness of breath, cough, and cyanogens with the respective impact rates of 0.87, 0.80, and 0.76. The impact rates of influential factors leading to the disease were found as drug addiction, family history, and exposure to allergic agents, Table 4.

Table 4. The most important symptoms and factors of COPD

COPD symptoms	**Impact rate**	**COPD factors**	**Impact rate**
Cyanogens	0.76	Drug addiction	0.78
Chest pain	0.78	Exposure to allergic agents	0.64
Shortness of breath	0.87	Family history	0.80
Phlegm production	0.76		
Cough	0.80		
Continuous cough cold	0.74		

Table 5. The generated rules from decision tree in diagnosis of COPD

Patient condition	Patient status
Smoking ≥ 5, family history ≥ 5, and chest pain ≥ 5	COPD
Smoking ≥ 5, family history ≥ 5, chest pain ≤ 5, and cyanogens ≥ 5	COPD
Smoking ≥ 5, family history ≥ 5, chest pain ≥ 5, and cyanogens ≥ 5, shortness of breath ≥ 5	COPD
Smoking ≥ 5, family history ≥ 5, chest pain ≥ 5, cyanogens ≤ 5, and shortness of breath ≤ 5	No COPD
Smoking ≥ 5, family history ≤ 5, chest pain ≥ 5, and shortness of breath ≥ 5	COPD
Smoking ≥ 5, family history ≤ 5, chest pain ≥ 5, and shortness of breath ≤ 5	No COPD
Smoking ≥ 5, family history ≥ 5, and chest pain ≤ 5	No COPD
Smoking ≤ 5, allergic environment ≥ 5, cough ≥ 5, and chest pain ≥ 5	COPD
Smoking ≤ 5, allergic environment ≥ 5, cough ≥ 5, chest pain ≤ 5, and shortness of breath ≥ 5	COPD
Smoking ≤ 5, allergic environment ≥ 5, cough ≥ 5, chest pain ≤ 5, and shortness of breath ≤ 5	No COPD
Smoking ≥ 5, family history ≥ 5, and cough ≤ 5	No COPD
Smoking ≥ 5 and allergic environment ≤ 5	No COPD

This algorithm also generated rules by which the health condition of each person can be predicted. These factors were also confirmed by Doctors' interviews and expert physicians. Figure 2 shows the decision tree model for prediction of the health condition of each individual by the disease factors and symptoms.

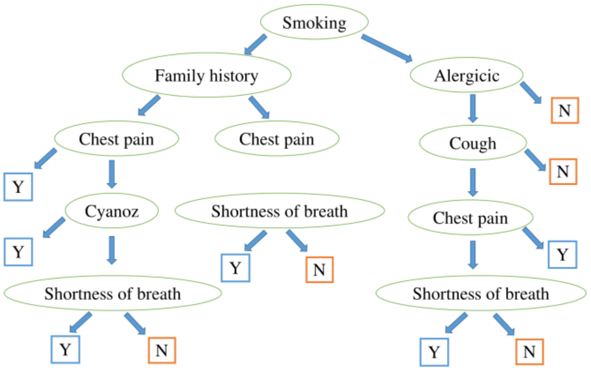

Figure 2. The decision tree for asthma disease prediction.

The outcomes of the decision tree were presented in Table 5. According to this rule set, the patients in risk of COPD were recognized with a symptom or factor with a level more than 5 days a week. These rules allow patient's self-management of COPD.

Discussion

One important subject in healthcare is the conversion of raw clinical data to meaningful information [10], mainly because there is not enough knowledge in many health care institutes while there have access to massive data [11]. The creation of information and excavation of patterns from abundant clinical data became of paramount importance in recent decade. Data mining techniques can be applied in this regard to generate useful information for disease diagnosis. It can help physicians to reach a consensus diagnosis. The patients subject to acute disease conditions can be detected in the early stages and appropriate medical treatments can be done [12-14].

Sanchez-Morillo et al., [15] was also applied the same data mining techniques to predict the disease status of patients subject to COPD. Dirven et al., [16] used the C4.5 decision tree algorithm (an older version of the C5 algorithm) to develop a model for rapid diagnosis of COPD. Based on their results, the smoking was the main contributor of the disease and the severe coughs was found as the main symptom. These findings were consistent with the findings of the present research. In the present work, the patients with acute conditions have symptoms such as shortness of breath, severe and painful coughs, and also phlegm production. These results are also consistent with the reports of physicians [17, 18]. According to the findings of the present research, about 80% of patients were smokers, consistent with the previous report in which smoking was introduced as the main risk factor for COPD [19]. Besides, the shortness of breath was found as the most influential factor with an impact rate of 0.87 that is consistent with the findings of Cristobalina et al. [5, 20]. They applied five data mining techniques and reached to the conclusion that dyspnoea is the best discriminating factor between people with and without COPD.

These results show that the design of standard questionnaire can lead us to correct diagnosis of the disease and subsequent spirometry prescription. This model is simple enough to be used by the personnel of clinics and hospitals as well [21].

The results of this model showed that the binary decision tree can guide us to more details of the rules. Accordingly, this tree works better than non binary ones. The high details of the created rules and high precision of accuracy and sensitivity confirms the right selection of the algorithm for the work.

Acknowledgement

The authors wish to acknowledge the financial support of Tarbiat Modares University (Tehran, Iran).

References

1. Qaseem, A., Wilt, T. J., weinberger, S. E., et al., Diagnosis and management of stable chronic obstructive pulmonary disease: A clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society Annals of Internal Medicine 2011. 155(3): p. 179-191.

2. Miravitlles, M., Calle, M., Soler-Cataluna, J. J., Clinical phenotypes of COPD: Identification, definition, and implications for guidelines Archivos de Bronconeumologia 2012. 48(3): p. 86-98.

3. Tan, J., Medical informatics: Concepts, methodologies, tools, and applications. 2009, New York: IGI Global.

4. Cano, I., Tenyi, A., Schueller, C., Wolff, M., Miguelanez, M. M. H., Gomez-Cabrero, D., Antczak, P., Roca, J., Cascante, M., Falciani, F., Maier, D. , The COPD knowledge base: Enabling data analysis and computational simulation in translational COPD research. Journal of Translational Medicine 2014. 12: p. 2-9.

5. Cristobalina, R.-A., Felix, R., Enrique, G.-D., Isidro, G.-M., Beatriz, C., Angeles, A., Real-data comparison of data mining methods in early detection of chronic obstructive pulmonary disease (COPD) in general practice Journal of Family Medicine and Disease Prevention 2016. 2(4): p. 1-7.

6. Sanchez-Morillo, D., Fernandez-Granero, M. A., Leon-Jimenez, A. , Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review Chronic Respiratory Disease 2016. 13(3): p. 264-283.

7. Al Jarullah, A.A., Decision tree discovery for the diagnosis of type II diabetes in International Conference on Innovations in Information Technology. 2011, IEEE: Abu Dhabi, United Arab Emirates

8. Alizadeh, S., Ghazanfari, M., Teimorpour, B., Data mining and knowledge discovery. 2011, Tehran, Iran: Publication of Iran University of Science and Technology.

9. Han, J., Kamber, M., Classification and prediction Data Mining: Concepts and Techniques. 2006, The Netherland Elsevier

10. Al Jarullah, A.A., Decision tree discovery for the diagnosis of type II diabetes International Conference on Innovations in Information Technology, 2011: p. 303-307.

11. Zhu, L., Wu, B., Cao, C., Introduction to medical data mining Journal of Biomedical Engineering, 2003. 20(3): p. 559-562.

12. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hermandez, K., Gopidi, R., Chang, J.-F., Hua, L., Data mining in healthcare and biomedicine: A survey of the literature. Journal of Medical Systems, 2012. 36(4): p. 2431-2448.

13. Tapak, L., Mahjub, H., Hamidi, O., Poorolajal, J., Real data comparison of data mining methods in prediction of diabetes in Iran. Healthcare Informatics Research, 2013. 19: p. 177-185.

14. Lee, B.J., Kim, J. Y., A comparison of the predictive power of anthropometric indices for hypertension and hypotension risk. Plos One, 2014. 9: p. e84897.

15. Sanchez-Morillo, D., Fernandez-Granero, M. A., Leon-Jimenez, A., Use of predictive algorithms in home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review. Chronic Repiratory Disease 2016. 13: p. 264-283.

16. Dirven, J.A.M., Tange, H. J., Muris, J. W. M., van Haaren, K. M. A., Vink, G., van Schayck, O. C. P. , Early detection of COPD in general practice: patient or practice managed? A randomized controlled trial of two strategies in different socioeconomic environments Primary Care Respiratory Journal 2013. 22: p. 331-337.

17. Zeng, X., Danquah, M. K., Chen, X. D., Lu, Y., Microalgae bioengineering: From CO2 fixation to biofuel production. Renewable and Sustainable Energy Reviews, 2011. 15: p. 3252-3260.

18. Price, D.B., Tinkelman, D. G., Halbert, R., Nordyke, R. J., Isonaka, S., Nonikov, D., Juniper, E. F., Freeman, D., Hausen, T., Levy, M. L., Qsterm, A., van der Molen, T., van Schayck, C. P., Symptom based questionnaire for identifying COPD in smokers. Respiration, 2006. 73: p. 285-295.

19. Burkhardt, R., Pankow, W., The diagnosis of chronic obstructive pulmonary disease Dtsch Arztebl International 2014. 111(49): p. 834-846.

20. Seyyed Shamsadin Athari, Seyyede Masoume Athari, Fateme Beyzay, Masoud Movassaghi, Esmaeil Mortaz, Mehdi Taghavi. Critical role of Toll-like receptors in pathophysiology of allergic asthma. European Journal of Pharmacology 2017; 808:21–27

21. Austin, P.C., A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines fr predicting AMI mortality Statistics in Medicine 2006. 26(15): p. 2937-2957.