



Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

REVUE FRANÇAISE
D'**Allergologie**

Revue française d'allergologie xxx (2019) xxx–xxx

Original article

A data-mining algorithm to assess key factors in asthma diagnosis

Un algorithme fondé sur les données pour l'évaluation des éléments clés du diagnostic de l'asthme

M. Mozaffarinya^a, A. Reza Shahriyari^a, M. Karim Bahadori^a, A. Ghazvini^a,
S. Shamsadin Athari^{b,*}, G. Vahedi^a

^a Baqiyatallah University of Medical Sciences, Tehran, Iran

^b Department of Immunology, School of Medicine, Zanjan University of Medical Sciences, Zanjan, Iran

Received 25 October 2018; accepted 24 January 2019

Abstract

Background. – Data mining is a technique widely used in medicine for the medical diagnosis, treatment and prognosis. Amongst the medical applications of data mining techniques, the most important are the construction of models comprising the key factors in disease diagnosis, evaluation of disease-related data patterns, monitoring of patient health status, and forecasting patients' future status. The purpose of this study was to assess key factors in the onset/progression of asthma central to accurate diagnosis of the disease.

Methods. – The survey was designed as a cross-sectional descriptive study in the summer of 2017. We selected patients diagnosed with asthma ($n = 300$) and individuals without asthma ($n = 80$) referred to clinics and hospitals in Tehran (Iran), Mofid Hospital (Tehran) and the Children's Medical Center (Tehran). A questionnaire was used to collect complementary medical information from the selected individuals. The recorded information was then evaluated and 200 of these cases were selected for further analysis. The surveyed data was analyzed using MATLAB and RapidMiner software and the results were assessed using a decision-tree algorithm.

Results and Discussion. – The key disease factors and their interrelation were analysed in the present study. Shortness of breath, wheezing, repeated respiratory attacks, cough, sleep disorder and facial bruising are common symptoms of asthma from the mildest to the most severe forms. Family history was also found to be an important factor in asthma risk for children. The findings of the study are consistent with those of expert physicians, with confirmation of the ability of data mining to generate knowledge from experimental data and to establish a consensual diagnosis.

© 2019 Elsevier Masson SAS. All rights reserved.

Keywords: Asthma; Data mining; Decision-tree algorithm; Medical diagnosis; RapidMiner

Résumé

Contexte. – L'exploration des données est une technique largement utilisée en médecine pour évaluer le diagnostic, le traitement et le pronostic d'une maladie. Parmi les applications médicales des techniques d'exploration de données, les plus importantes sont la construction de modèles comprenant les éléments clés du diagnostic d'une affection, l'évaluation des schémas de données associées à la maladie, la surveillance de l'état de santé des patients et la prévision de leur état futur. Le but de cette étude était d'évaluer les facteurs clés dans l'apparition et la progression de l'asthme, essentiels au diagnostic précis de la maladie.

Méthodes. – Cette enquête est une étude descriptive transversale réalisée pendant l'été 2017. Nous avons sélectionné des patients asthmatiques chez lesquels le diagnostic d'asthme avait été porté ($n = 300$) et des personnes non asthmatiques ($n = 80$) adressées à des cliniques et des hôpitaux de Téhéran (Iran), à l'hôpital Mofid (Téhéran) et au centre médical pour enfants (Téhéran). Un questionnaire a été utilisé pour collecter des informations médicales complémentaires auprès des individus sélectionnés. Les informations enregistrées ont ensuite été évaluées et 200 de ces cas ont été sélectionnés pour une analyse plus approfondie. Les données de l'enquête ont été analysées à l'aide des logiciels MATLAB et RapidMiner et les résultats ont été évalués à l'aide d'un algorithme décisionnel.

* Corresponding author.

E-mail address: SS.Athari@gmail.com (S. Shamsadin Athari).

Résultats et discussion

Les facteurs clés du diagnostic d'asthme et leurs relations réciproques ont été analysés dans la présente étude. Essoufflement, respiration sifflante, événements respiratoires à répétition, toux, troubles du sommeil et aspect bleuté du visage sont des symptômes courants de l'asthme, de la forme la plus légère à la forme la plus grave. Les antécédents familiaux étaient également un facteur important du risque d'asthme chez les enfants. Les résultats de l'étude concordent avec ceux des médecins experts, confirmant ainsi la capacité de l'exploration de données à générer des connaissances à partir de données expérimentales et pour établir un diagnostic consensuel.

© 2019 Elsevier Masson SAS. Tous droits réservés.

Mots clés : Asthme ; Exploration de données ; Algorithme décisionnel ; Diagnostic médical ; Logiciel RapidMiner

1. Introduction

Asthma comes from the Greek word  “*ásthma*”, meaning “panting”. It is a long-term inflammatory disease of the airways of the lungs characterized by variable and recurring symptoms, reversible airflow obstruction and bronchospasm [1]. During acute asthma attacks, symptoms include shortness of breath, wheezing and coughing [1,2]. Depending on the individual, these symptoms occur every day in some patients [2]. Control and treatment of the disease turns on relieving stress psychologically and by drugs. However, the significant progress made in medical science has not yet enabled complete success in asthma treatment. Besides, the drugs commonly prescribed for the condition have certain side effects [3,4].

Numbers of asthma patients have increased significantly since 1970, reaching 300 million people in 2010. In 2009, the disease caused 250 deaths worldwide. The problem is exacerbated by the shortage of expert physicians in the developing world. Since early detection of the disease undoubtedly leads to effective treatment, it is essential to design an algorithm enabling accurate rapid diagnosis of the key factors affecting asthma severity.

Data-mining techniques are useful tools for identifying patterns from data and designing predictive models [5]. They can identify novel, accurate, and intelligible patterns between data to uncover the knowledge lying behind massive data. These techniques have been applied in healthcare to gain in-depth understanding of medical data, to select appropriate therapeutic strategies following accurate diagnosis, and to choose suitable medical care based on disease prognosis [6,7]. The various methods, which differ in precision and accuracy, have been applied by different researchers. For example, the C5 decision-tree algorithm was used to analyze microarray data and identify the genes involved in glaucoma [8]. Sanchez-Morillo et al. [9] used data mining techniques to establish the prognosis for patients with persistent lung disease and in the remote monitoring of their respiratory systems. The authors processed and classified data from asthma patients using data mining. They found that asthma is not age-related and that the specific conditions of individuals can affect their asthma symptoms. Bereznicki et al. [10] investigated the history of drug consumption in asthma patients in order to manage this usage. Hans et al. [11] applied data mining techniques to classify data from asthma patients. Wang et al.

Table 1
Demographic features of participants.

Disease	Demographic features							
	Age group (years)			Sex		Statistics		
	6–11	12–39	40+	Male	Female	Mean	SD	Mode
Asthma	103	28	20	78	73	15.8	14.8	6
Healthy	13	21	29	30	34	35.3	19.5	10

[12] used the C5 decision tree to develop a model for asthma management in children.

Given the significance of asthma in healthcare, it is a key priority to develop an optimal method for its accurate and rapid medical diagnosis. The cultural, religious, racial, and climatic conditions of Iran differ from those of Europe, North America and many other places. These differences underscore the need to develop a new model for diagnosis of the disease in suspected candidates living in countries comparable to Iran. The purpose of the present study is to develop such a model in order to understand the patterns behind the key factors for asthma diagnosis. These key factors were collated from previous studies.

2. Methodology

2.1. Participants

A population of 380 individuals diagnosed with asthma by physicians (and confirmed by laboratory procedures) and referred to clinics across Tehran (Iran) was surveyed using a questionnaire. Asthma was diagnosed with the help of both standard medical diagnostic guidelines (e.g. GINA, NAEPP and VA/DoD) and routine laboratory procedures in the clinics/hospitals. Questionnaires with missing information or erroneous data were omitted from the study. Thus, out of a total of 380 questionnaire files collected, 214 cases were selected for further analysis. Among the selected cases, 151 subjects had been diagnosed with asthma while 63 individuals had temporary or mild pulmonary symptoms unrelated to asthma and were therefore considered healthy in pulmonary terms. The demographic characteristics of participants are shown in Table 1. The questionnaire comprised 42 questions on general background (e.g. age, profession, etc.) and clinical status (e.g. cough, abscess, pain, etc.).

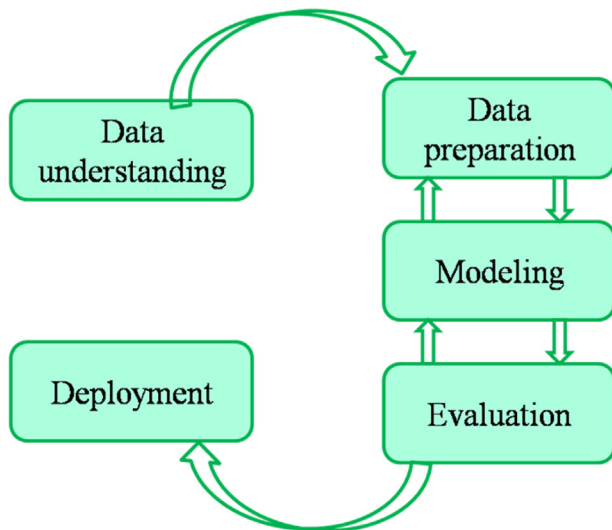


Fig. 1. Process diagram showing the five phases of CRISP data mining.

2.2. The road map

The decision tree is a common data mining technique. It has a flowchart-like structure in which the highest node represents the root and each branch is the outcome of the tests. The leaf nodes represent classification rules. The rules generated by analysis are expressed as “if” and “then” [11]. The C5 decision-tree algorithm was used in the RapidMiner environment to excavate the knowledge behind the data gathered for asthma and to develop a pattern for diagnosis of the condition.

The gathered data were first analyzed in the MATLAB environment and then classified by degree of importance using the SMC function. This information was subsequently used to design a predictive model to diagnose each patient’s condition and the impact of each factor. The CRISP method was used in the algorithm design [12]. The CRISP method comprises five phases as shown in Fig. 1. These phases are discussed in the following sections.

2.3. Understanding data

Understanding data is the first step in the design of each system as well as in determining data mining goals and success criteria. Given the lack of successful treatment for the rising tide of asthma, as well as the similarity of symptoms between diseases of different kinds, it is hoped that the analysis of data gathered from our patients will prove useful for others.

2.4. Data preparation

Patients were surveyed by means of a questionnaire in 2017 to gather pertinent information. The data were then quality tested and checked. To this end, the case study was carried out by considering the files of patients referred to the Mofid Hospital and the Children’s Medical Center. The characteristics used concerned personal and family history, life style and disease symptoms. These study characteristics are presented in Table 2

in the parameter column. The first column shows the study characteristics in the data set while the second and third columns show the parameter type and the impact of each factor on asthma, respectively. To optimize the raw data, the categorical data was converted to numerical data and incomplete data was removed. Any slight inconsistency between data was corrected by defining a preprocessing block diagram in RapidMiner during the modeling procedure.

2.4.1. Reliability of the questionnaire

Reliability is the extent to which a questionnaire produces the same results in repeated trials. Cronbach’s alpha was used to measure reliability. Values of between 0.7 and 0.9 indicate good consistency. The value for the present study was 0.859, thus confirming the reliability of the questionnaire.

2.4.2. Validity of the questionnaire

Validity indicates how close the measurements are to the actual values. It can be evaluated by different means based on the knowledge of experts and instructions from standard asthma evaluation guidelines such as GINA, NAEPP and VA/DoD. To this end, comments were obtained from 11 expert physicians and clinical symptoms known to play a role in asthma were retrieved from standard guidelines. The extracted information was then used in finalizing the questionnaire.

2.5. Modeling

At first, the impact of influential factors and symptoms on asthma was considered using the SMC method in a MATLAB environment. Modeling was done in RapidMiner. The C5 decision-tree algorithm was used to build either a decision tree or a rule set. The decision tree explicitly describes branching by use of the algorithm so that the final leaf node of the graph describes a subset of training data while each sample in the training section belongs to a final node in the tree. For the C5 tree designed in this study, the inputs of the algorithm are all patient characteristics and the outcomes are patient status. The original data set is split into two training and test sets with respective percentages of 30% and 70%. The models were generated using the 10-split method such that the data set was randomly divided into 10 parts [8]. Subsequently, each of these 10 parts was chosen as the test data and the nine others as training data. This strategy was chosen mainly because it provides the most accurate predictions. The class labels (Y for asthmatic and N for non-asthmatic) were chosen by calculating precision (accuracy). Then, using the SMC method, the main factors concerning disease onset and severity were chosen as the tree branches. The branching divisions were based on the Gini index, with each tree being split into sub-branches to include the data for each node within a class. The leaf-to-root path was then followed in the reverse direction and the rules generated are expressed conditionally.

2.6. Evaluation

Evaluation of the model results is essential to determine whether it is suitable for use in its current form or if it requires

Table 2
Characteristics of the asthma disease data set and results of analysis of the impact of each data item on asthma.

Parameter	Categorical Type	Impact rate	Parameter	Categorical type	Impact rate
Age	Numerical		Sweating	Nominal	0.32
Gender	Nominal		Fever	Nominal	0.28
Job	Nominal		Weight loss	Nominal	0.49
Length	Numerical		Anesthesia	Nominal	0.34
Weight	Numerical		Antibiotic consumption	Nominal	0.28
Shortness of breath	Nominal	0.87	Pollutant presence	Nominal	0.56
Wheezing chest	Nominal	0.74	Inability to talk	Nominal	0.34
Dry cough	Nominal	0.78	Aspirin usage	Nominal	0.34
Atopic disease(s)	Nominal	0.42	Smoker pregnant mother	Nominal	0.29
Family history	Nominal	0.80	Drug addiction	Nominal	0.26
Exposure to allergens	Nominal	0.64	Cough, cold	Nominal	0.35
Bruised nails and lips	Nominal	0.76	Immunodeficiency disease	Nominal	0.24
Restlessness	Nominal	0.67	Disease awareness	Nominal	0.77
Stress	Nominal	0.54	Chest pain	Nominal	0.68
Depression	Nominal	0.68	Strong respiratory attacks	Nominal	0.76
Stomach reflux	Nominal	0.10	Repeated attacks	Nominal	0.72
Work absent	Nominal	0.33	Emergency need	Nominal	0.46
Sleep disorder	Nominal	0.78	Hospitalization	Nominal	0.53
Wheezing with activity	Nominal	0.80	Reception in priority	Nominal	0.41
Bad breath	Nominal	0.25	Cough with sputum	Nominal	0.30
			Cyanogen	Nominal	0.76

Bold entries indicate the most significant symptoms and asthma factors.

Table 3
Perturbation matrix for evaluation of model precision.

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

TN: number of records of negative class for which the classification algorithm correctly assigned class as negative; TP: records of positive class for which the classification algorithm correctly assigned class as positive; FN: records of positive class for which the classification algorithm assigned class as negative; FP: records of negative class for which the classification algorithm assigned class as positive.

further improvement. To evaluate the accuracy of the model, the data were divided into three sets: training, test, and validation. The training set was used to design the decision tree and the test data set was used for tree evaluation and assigning a class label. The validation set was used to test correct classification in the model. There are various indexes for evaluating correct classification of the results by the classification methods, such as allergy, transparency, accuracy and precision, etc. [13,14]. The accuracy of each classification of the training data set is the percentage of training set observations correctly classified by the method. The test data were used to calculate this index. The precision of the model was evaluated using the perturbation matrix (Table 3). If data is set out in class M, a class matrix has a minimum size of $M \times M$. The ideal case is that having the most data relevant to observations on the main diagonal of the matrix with the remainder of the matrix values near or equal to zero.

The precision or rate-of-classification algorithm is the best performance metric that measures the total accuracy of a classification. This criterion is the commonly used metric for evaluation of the performance of classification algorithms. It indicates the

extent to which the designed classification correctly predicts the test records.

Classification precision may be calculated using the following equation:

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

where TP and TN are the main values to be optimized in a two-class problem.

2.7. Deployment

Following design and validation of the model, it can be deployed in two ways. First, the analyst uses the model and the results obtained to draw conclusions from the data. Second, the model can be used either for assigning records based on their classifications or else weighting them for special treatment. The results of data mining can be combined using the knowledge of data experts. In a diagnostic system, existing patterns of a process can be combined with newly discovered patterns.

3. Results

In the present study, all key factors concerning asthma, including personal information, family history, life style, and clinical information on the disease, were considered. The impact of key factors on the disease is highlighted in Table 2 using bold type. Subsequently, using the decision-tree algorithm, the most important symptoms of the disease were found to be shortness of the breath, wheezing on activity, cyanogens, strong respiratory attacks, and dry cough, with respective impact rates of 0.87, 0.80, 0.78, 0.76, and 0.76. Family history was found to be the most influential factor for the disease, with an impact rate of

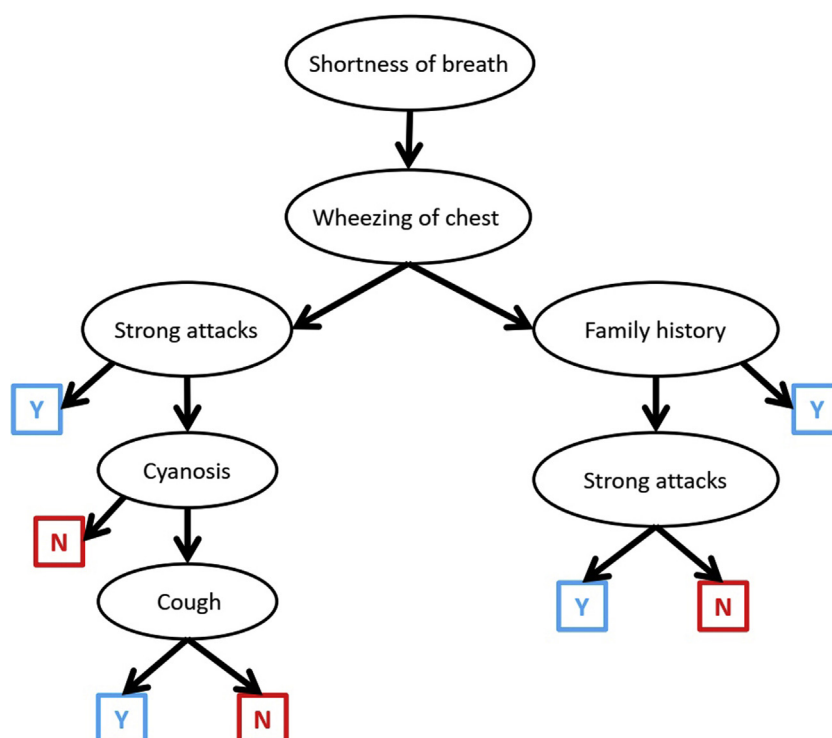


Fig. 2. Decision tree for asthma prediction.

Table 4

Rules for asthma diagnosis generated using the decision tree.

Patient condition	Patient status
Shortness of breath ≥ 5 , wheezing on activity ≥ 5 , and family history ≥ 5	Asthma
Shortness of breath ≥ 5 , wheezing on activity ≥ 5 , family history ≤ 5 , and strong attacks ≥ 5	Asthma
Shortness of breath ≤ 5 , wheezing on activity ≤ 5 , and strong attacks ≥ 5	No asthma
Shortness of breath ≥ 5 , wheezing on activity ≥ 5 , strong attacks ≤ 5 , cyanogens ≥ 5 , and cough ≥ 5	Asthma
Shortness of breath ≤ 5 , wheezing on activity ≥ 5 , strong attacks ≤ 5 , cyanogens ≥ 5 , and cough ≤ 5	No asthma
Shortness of breath ≤ 5 , wheezing on activity ≥ 5 , strong attack ≤ 5 , and cyanogens ≤ 5	No asthma

0.80. This algorithm also generated rules by which the health condition of each individual may be predicted. These factors were also confirmed by the expert physicians. Fig. 2 shows the decision-tree model for prediction of the health status of individuals based on disease factors and symptoms. The accuracy of the algorithm in prediction of asthma and classification of patients based on the presence/absence of asthma as true Y and true N was 98.68% (~99%). When predicting Ys, the predicted true Ys totalled 149 cases while the predicted true Ns totalled 2. Thus, using the data for asthma patients, the model was able to predict their condition with ~99% accuracy. Of all the patients referred to the clinics/hospital, 149 individuals had asthma. Asthma was correctly diagnosed for all patients except two. Thus, the accuracy of asthma diagnosis was 97.63%. The remaining 51 cases were referred to clinics/hospital for other diseases and all were correctly diagnosed.

The outcomes of the decision tree are presented in Table 4. Patients at risk for asthma were identified with a symptom or factor with a significant level for more than 5 days a week. These rules enable self-management of their asthma by patients.

4. Discussion

Asthma and allergy are chronic diseases caused by genetic, medical, and environmental factors. The complexity of these diseases on the one hand and their misdiagnosis on the other lead to difficulties in proper diagnosis and control of the disease. This issue is clearly felt in developing countries, where the shortage of expert physicians is a problem. Knowledge generated from patient records using data mining techniques can be useful for effective diagnosis of the disease. Indeed, the objective of data mining is to generate knowledge from recorded data in patient files and databases to develop intelligible patterns. To excavate information from the records, a new questionnaire was designed in this study. However, a number of relevant questionnaires have been published by other researchers for their own case studies [15–17]. Data mining algorithms such as the C5 decision tree are useful tools to assist physicians in medical diagnosis. In the creation of smart systems, this knowledge is often unreliable, for two possible reasons: (1) the knowledge of physicians cannot quantitatively relate the key disease conditions to potential

outcomes; (2) the results of experimental studies may not be correct.

According to the results of the present study, patients with poor health status presented acute shortness of breath and also exhibited symptoms such as wheezing and acute cough. The data mining results are consistent with the results of previous medical studies. According to the findings of the model proposed in the present study, eighty percent of patients had a family history of asthma [18,19]. This finding is consistent with the physicians' reports obtained during interviews with doctors [20]. It was shown that asthma is not age-related and this finding is consistent with the reports by Hans et al. [11], which state that asthma is the most rapidly increasing disease in which age is not a key factor. As indicated in other reports [21], asthma is not sensitive to pharmaceutical treatment. Antibiotic and aspirin usage respectively have impact rates of only 0.28 and 0.34 (Table 2).

In this study, the inputs were data sets consisting of values of either zero or one indicating respectively healthy or disease status. The system outcome is a decision-tree algorithm.

The results of this system were based on disease symptoms and family history. From 42 factors identified as key factors or symptoms, seven were assigned as central factors for asthma. The designed system can be successfully used by individuals for self-diagnosis of their disease.

Disclosure of interest

The authors declare that they have no competing interest.

Acknowledgements

The authors would like to give their sincere thanks to Tarbiat Modares University (Tehran, Iran) for providing support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.reval.2019.01.013>.

References

- [1] Athari SM, Afshari F, Eftekhari A, Athari SS. The surveillance system, diagnosis and treatment challenges of asthma and health policy orientation of main challenges. *J Pain Manag Ther* 2017;1(2):1–5.
- [2] Wenzel SE. Complex phenotypes in asthma: current definitions. *Pulm Pharmacol Ther* 2013;26:710–5.
- [3] Sardesai A, Kharat V, Sambarey P, Deshpande A. Efficiency of fuzzy-stat modelling in classification of gynaecologists and patients. *J Intell Syst* 2016;25:147–57.
- [4] Czajkowski M, Kretowski M. Top scoring pair decision tree for gene expression data analysis in Software tools and algorithms for biological systems. New York: Springer; 2011. p. 27–35.
- [5] Khajehei M, Etemady F. Data mining and medical research studies. Tuban, Indonesia: IEEE; 2010.
- [6] Jayalakshmi T, Santhakumaran A. A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. Bangalore, India: IEEE; 2010. p. 159–63.
- [7] Narooie-Najad M, Paylakhi SH, Shojaei S, Fazlali Z, Rezaei Kanavi M, Nilforushan N, et al. Loss of function mutations in the gene encoding latent transforming growth factor beta binding protein 2, LYBP2, cause primary congenital glaucoma. *Hum Mol Genet* 2009;18:3969–77.
- [8] Al Jarullah AA. Decision tree discovery for the diagnosis of type II diabetes. Abu Dhabi, United Arab Emirates: IEEE; 2011.
- [9] Sanchez-Morillo D, Fernandez-Granero MA, Leon-Jimenez A. Use of predictive algorithms in home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. *Chron Respiratory Dis* 2016;13:264–83.
- [10] Bereznicki BJ, Peterson GM, Jackson SL, Walters EH, Fitzmaurice KD, Gee PR. Data mining of medication records to improve asthma management. *Med J Aust* 2008;189:21–5.
- [11] Hans A, Kaur S, Singh N. Cloud based data mining model for asthma diagnosis. *Int J Grid Distributed Comput* 2016;9:317–26.
- [12] Wang T, Shao K, Chu Q, Ren Y, Mu Y, Qu L, et al. Automics: an integrated platform for NMR-based metabolomics spectral processing and data analysis. *BMC Bioinformatics* 2009;10:83.
- [13] Alizadeh S, Ghazanfari M, Teimorpour B. Data mining and knowledge discovery. Tehran, Iran: Publication of Iran University of Science and Technology; 2011.
- [14] Han J, Kamber M. Classification and prediction data mining: concepts and techniques. The Netherlands: Elsevier; 2006.
- [15] Toren K, Brisman J, Jarvholm B. Asthma and asthma like symptoms in adults assessed by questionnaires: a literature review. *Chest* 1993;104:600–8.
- [16] Jenkins MA, Clarke JR, Carlin JB, Robertson CF, Hopper JL, Dalton MF, et al. Validation of questionnaire and bronchial hyperresponsiveness against respiratory physician assessment in the diagnosis of asthma. *Int J Epidemiol* 1996;25(3):609–16.
- [17] Fernandez CM, Fernandez-Benitez M, Miranda MP, Grima FG. Validation of the Spanish version of the phase III ISAAC questionnaire on asthma. *J Investig Allergol Clin Immunol* 2005;15:201.
- [18] Louviere JJ, Woodworth G. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *J Mark Res* 1983;20(4):350–67.
- [19] Motschnig-Pitrik R. Analyzing the notions of attribute, aggregate, part and member in data/knowledge modelling. *J Sys Software* 1996;33:113–22.
- [20] Morice A. The diagnosis and management of chronic cough. *Eur Respir J* 2004;24:481–92.
- [21] 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. *J Fam Med Dis Prev* 2016;2:1–7.