JRHS 2015; 15(3): 189-195

JRHS
Journal of Research in Health Sciences
journal homepage: www.umsha.ac.ir/jrhs

Original article

Presentation of a model-based data mining to predict lung cancer

Reza Shahhoseini (MSc)[ae], Ali Ghazvini (MD)[b], Mansour Esmaeilpour (PhD)[c], Gholamhossein Pourtaghi (PhD)[d], Shahram Tofighi (PhD)[e*]

[a] *Department of Health Care Management, School of Health, Baqiyatallah University of Medical Sciences, Tehran, Iran*
[b] *Department of Internal Medicine, School of Medicine, Baqiyatallah University of Medical Sciences, Tehran, Iran*
[c] *Department of Computer Engineering, Hamadan Branch, Islamic Azad University, Hamadan, Iran*
[d] *Health Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran*
[e] *Health Management Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran*

ARTICLE INFORMATION

Article history:
Received: *11 April 2015*
Revised: *16 May 2015*
Accepted: *14 July 2015*
Available online: *08 September 2015*

Keywords:
Data Mining
Lung Cancer
Decision Tree
Neural Networks

**** Correspondence***
Shahram Tofighi (PhD)
Tel: *+98 912 394 66 48*
E-mail: *shr_tofighi@yahoo.com*

ABSTRACT

Background: The data related to patients often have very useful information that can help us to resolve a lot of problems and difficulties in different areas. This study was performed to present a model-based data mining to predict lung cancer in 2014.

Methods: In this exploratory and modeling study, information was collected by two methods: library and field methods. All gathered variables were in the format of form of data transferring from those affected by pulmonary problems (303 records) as well as 26 fields including clinical and environmental variables. The validity of form of data transferring was obtained via consensus and meeting group method using purposive sampling through several meetings among members of research group and lung group. The methodology used was based on classification and prediction method of data mining as well as the method of supervision with algorithms of classification and regression tree using Clementine 12 software.

Results: For clinical variables, model's precision was high in three parts of training, test and validation. For environmental variables, maximum precision of model in training part relevant to C&R algorithm was equal to 76%, in test part relevant to Neural Net algorithm was equal to 61%, and in validation part relevant to Neural Net algorithm was equal to 57%.

Conclusions: In clinical variables, C5.0, CHAID, C & R models were stable and suitable for detection of lung cancer. In addition, in environmental variables, C & R model was stable and suitable for detection of lung cancer. Variables such as pulmonary nodules, effusion of plural fluid, diameter of pulmonary nodules, and place of pulmonary nodules are very important variables that have the greatest impact on detection of lung cancer.

Citation: Shahhoseini R, Ghazvini A, Esmaeilpour M, Pourtaghi G, Tofighi S . Presentation of a model-based data mining to predict lung cancer. J Res Health Sci. 2015; 15(3): 189-195.

Introduction

Cancer has different types and is the second cause of death in the world. It is annually estimated that 30 million people lose their lives due to cancer. On the other hand, about 30% of total deaths due to cancer is preventable[1]. During the recent decades, despite the successes in the context of control and prevention of epidemic diseases, the level of prevalence and morbidity of chronic diseases has considerably increased[2]. According to the National Cancer Institute of America in 2014, the number of cancer cases and deaths from lung cancer was estimated 159260 deaths in both sexes[3]. In Iran, after the cardiac-vascular diseases and accidents, cancer is the third cause of death[4]; while, of every 6.7 million-death cases due to cancer, 1.18 million occurs among people with lung cancer[5].

In developed countries, the rate of lung cancer cases among men and women is respectively 61 and 18.9 per one hundred people. On the other hand, in developing countries, it is 28.7 and 10.4 for men and women respectively. Moreover, 39.9 cases of men and 13.5 cases of women have lung cancer in the whole world. Based on the 2002 report of Globocan in Iran, ASR of men and women is 8 and 2.4 correspondingly[6].

Furthermore, according to the information obtained from recorded cancer cases in Ardebil, ASR of lung cancer among men and women was respectively 7.9 and 3.59/100000 people[7]. The number of years lost because of death due to cancer arises to 650 thousands. The researchers of American Cancer Society stated that cancer produced a cost of 895 billion dollars for American economy in 2008 and this number is equivalent to 1.5% of internal gross production in the whole world[8].

People with certain jobs have better pre-awareness about cancer as well as its clinical and environmental signs compared to others; this superiority can be due to the social and economic situation of this group of patients[9]. Some in advance alerts in hard works, unfavorable working

conditions, long hours of work, noise in workplace, tiring physical works, continuous and/or repetitive muscular pressure, mental pressures and dissatisfaction with job have relationship with early spread of retirement among personnel[10].

In recent years, a progressive flow has been made in order to use the computational and computerized sciences in health care systems. The use of information technology, which has considerably resulted in improving efficiency and effectiveness of health cares, is counted as an important strategy in the context of health, because the use of these technologies is effective on the lives of all people[11].

Medical data mining has high potential for discovering the hidden models presented in data, and these models can be used for clinical diagnoses[12].

In addition, the warner models of prediction have high potential to help in decision-making process relevant to patient's conditions. There are multivariable models that use several warner variables to predict proper diagnosis and treatment for individuals, and, based on them, the health interventions or changing life style can be done about individuals who are subject to danger. In addition, the application of these models can be useful in identifying individuals who are prone to a special kind of disease and can be used to manage the treatment and recovery of disease outcomes[13].

The clinical decision-makings are often evidence-based and based on physician's experiment, whereas in databases, the data rich in knowledge have been neglected and this will result in undesirable orientations, medical errors, and inordinate increments of diagnostic and therapeutic costs. The presentation of a system is necessary to unify support of clinical decision-makings and computerized medical files to decrease medical errors, increment patient safety, reduce unnecessary surgical operations, improve patient caring, and to decrease health costs. Reaching these goals is possible via data modeling and analysis tools. Data mining has the possibility of making an environment rich with knowledge to enhance quality and support of clinical decision-making[14].

By identifying and studying the clinical and environmental factors relevant to diagnosis or death due to cancer, this study was aimed to offer a model-based data mining to predict lung cancer.

Methods

In this exploratory and modeling study, data were collected by two methods: library and field methods. The research community included the patients referring to Baqiyatallah Hospital, Tehran, Iran. This hospital's database is composed of 303 records[12]. Regarding that data mining is not dependent on sample size; therefore, the number of research samples was randomly selected among the individuals affected by pulmonary problems and had referred to Baqiyatallah therapeutic centers. Consequently, the database of this research included the records related to patients (150 patients), and the records related to people whose probability of having lung cancer was rejected after medical studies (153 persons). The variables of this research were clinical and environmental within 21 fields including age, sex, incidence of cancer among first degree relatives, smoking, quitting or continuing addiction, time of quitting smoking, occupational pollutants, shortness of breath, hemoptysis, existence of pulmonary nodules, place of pulmonary nodules, history of non-pulmonary diseases especially cancer, history of previous affection, cough, findings of effusion of pleura fluid, spirometry, hospital ward of patient admission, chest graph, geographical area in which they work, and lung cancer (Table 1).

Table 1: Syntax file of research variables

Type of Variable in Dataset	Variable's Type
Age	Numerical
Sex	Two-Valued
Incidence of cancer among first degree relatives	Classical
Smoking	Numerical
Addiction	Two-Valued
Time of quitting smoking	Numerical
Occupational pollutants	Classical
Shortness of breath	Classical
Hemoptysis	Numerical
Pulmonary nodules	Two-Valued
Diameter of pulmonary nodules	Classical
Place of lung nodules	Classical
History of previous non-pulmonary diseases	Classical
History of previous pulmonary diseases	Classical
Cough	Classical
Effusion of pleura fluid	Classical
Spirometry	Two-Valued
Patient acceptance	Classical
Chest graph	Two-Valued
Geographical area served	Classical

The methodology used in this study is based on method of classification and prediction of data mining as well as method of supervision with algorithms of CART, CHAID, C5.0 and neural networks in two groups of clinical and environmental variables and 8 algorithms using Clementine 12 software.

The inclusion criteria were as follows:

a. Form designing: This includes the extraction of data from the patients' medical record by research team and professional doctors (pulmonologists). The output of this step is the form of data extraction. In this step, the informational fields related to research goals were recognized via the scientific resources by participation of members of research group and lung group. The validity of the form of data transferring was obtained via consensus and meeting group methods and by using purposive sampling through several meetings among members of research group as well as lung group. We applied the Crisp's standard model alongside methods of classification and prediction of data mining. The classification was performed by method of supervision with four algorithms of CART, CHAID, C5.0 and neural networks in two groups of clinical and environmental variables and 8 algorithms by the use of Clementine 12 software.

b. Study of records: The extraction of data from the records with confirmation of professional doctors.

c. Designing computerized form of data entrance: In this step, after designing and confirming the validity of the form of data transferring, the extracted data were codified

for every person in the format of birth certificate and, therefore, the database (data store) was formed.

d. Information entrance: The data recorded in the form of data transferring were transferred to Excel software and the data store or database was inserted into the environment of Clementine software.

e. The extracted data were preprocessed during data cleaning (imperfect data, data accompanied by error, disagreeable and repetitive data).

f. Data integrating: The data are converted from such different resources as records, databases etc. to a realizable structure for model.

g. Data limiting: The data with different ranges must be converted to the classified ranges.

h. Discrete-making: The classified data are converted to the classified ranges (three states of 0,1 and 2)

i. Data converting: The conversion of data for entering to the learner standard models.

j. Based on the research goals, the proper questions were designed to determine the relationship between the data components (variables) by using the learner mechanized models. A number of models are outputs of this step.

k. Confirming model by lung experts panel, which the output of this step is confirmed models, and the performance of determined models was investigated via ROC graph.

l. The specific preventive policies were adjusted according to the type of variables affecting the creation of lung cancer and their influence intensity that was obtained from the models in the format of rules. The accuracy of the four algorithms has been based on the following formula:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false pesitives} + \text{false negatives} + \text{true negatives}}$$

Results

The findings of this research were obtained through two parts of modeling with clinical and environmental variables. In this model, data are divided into three sections: training, test, and validation.

First section

Modeling with clinical variables: In the first four models, the input data included age, sex, incidence of cancer among first degree relatives, shortness of breath, hemoptysis, pulmonary nodules, diameter of pulmonary nodules, place of lung nodules, history of non-pulmonary diseases, history of previous pulmonary diseases, cough, pleural effusion, spirometry, patient acceptance as well as chest graph. Here the target is cancer.

Precision of first model: Table 2 shows the C5.0 algorithm. As indicated in table 2, the precision of the model is high in all three parts; therefore, this model is a steady and reliable model.

Table 2: Model's precision in training for output field cancer, test and validation (clinical variables)

Partition	Training	Percent	Testing	Percent	Validation	Percent
Comparing $C-Cancer with Cancer (The C5.0 algorithm)						
Correct	174	100	80	100	49	100
Wrong	0	0.0	0	0.0	0	0.0
Total	174	100	80	100	49	100
Comparing $R-Cancer with Cancer (The CHAID algorithm)						
Correct	174	100	80	100	49	100
Wrong	0	0.0	0	0.0	0	0.0
Total	174	100	80	100	49	100
Comparing $R-Cancer with Cancer (The C&R algorithm)						
Correct	174	100	80	100	49	100
Wrong	0	0.0	0	0.0	0	0.0
Total	174	100	80	100	49	100
Comparing $N-Cancer with Cancer (The Neural Net algorithm)						
Correct	174	100	79	98.7	49	100
Wrong	0	0.0	1	1.3	0	0.0
Total	174	100	80	100	49	100

Precision of second model: The CHAID algorithm is according to Table 2. The precision of the model is high in all three parts, thus this model is a steady and reliable model.

Precision of third model: The C&R algorithm is according to Table 2. The precision of this model is high in all three parts, so this model is a steady and reliable model.

Precision of fourth model: The Neural Net algorithm is shown in Table 2. The precision of model is high in all three parts, and this model is a steady and reliable model as well.

Findings from evaluating models' precision

The statistical analyzers expect that the models having high precision be close to the best line presented in the figure and, in fact, be bowl-shaped. In this figure, the horizontal axis is disjunction points and the vertical axis is cumulative percentage of samples, which are located under these disjunction points. On the other hand, the linear model closer to the best line, namely "BEST", is a better model (Figure 1). Meanwhile, models' precision with clinical variables for training, test, and validation data are shown in Table 3.

Table 3: Model's precision in training, test and validation (clinical variables)

Model's name	Model's precision in training %	Model's precision in test %	Precision of validation %
C5.0	100	100	100
CHAID	100	100	100
C & R	100	100	100
Neural Net	98.8	100	100

Second section

Modeling with environmental variables: In the second four models, the input data are smoking, addiction, time of quitting smoking, occupational pollutants, and geographical area in which they work, and the target is cancer.

Precision of fifth model: The C5.0 algorithm is according to Table 4. The model's precision in training, test, and validation sections is 65%, 60% and 55% respectively.

Precision of sixth model: The CHAID algorithm is according to table 4. The model's precision in training, test, and validation sections is 65%, 57% and 55% correspondingly.

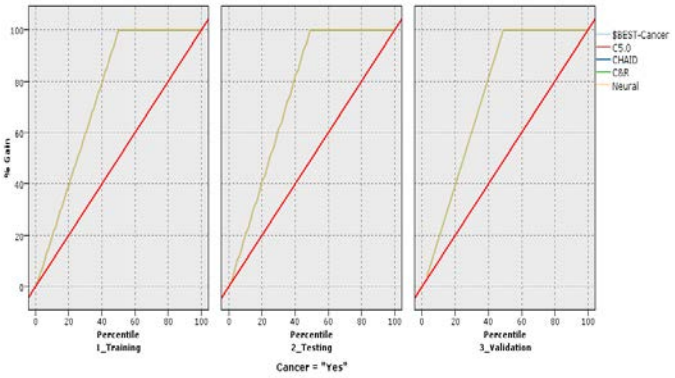

Figure 1: ROC figure of clinical variables

Table 4: Model's precision in training for output field cancer, test and validation (environmental variables)

Partition	Training	Percent	Testing	Percent	Validation	Percent
Comparing $C-Cancer with Cancer (The C5.0 algorithm)						
Correct	113	64.9	48	60.0	27	55.1
Wrong	61	35.1	32	40.0	22	44.9
Total	174	100	80	100	49	100
Comparing $R-Cancer with Cancer (The CHAID algorithm)						
Correct	113	64.9	46	57.5	27	55.1
Wrong	61	35.1	34	42.5	22	44.9
Total	174	100	80	100	49	100
Comparing $R-Cancer with Cancer (The C&R algorithm)						
Correct	132	75.9	43	53.7	28	57.1
Wrong	42	24.1	37	46.3	21	42.9
Total	174	100	80	100	49	100
Comparing $N-Cancer with Cancer (The Neural Net algorithm)						
Correct	106	60.9	49	61.3	28	57.1
Wrong	68	39.1	31	38.7	21	42.9
Total	174	100	80	100	49	100

Precision of seventh model: The C&R algorithm is according to Table 4. The model's precision in training, test, and validation sections is 76%, 54% and 57% in that order.

Precision of eighth model: The Neural Net algorithm is according to Table 4. The model's precision in training, test, and validation sections is 61%, 61% and 57% respectively.

Findings from evaluating models' precision

The statistical analyzers expect that the models having high precision be close to the best line presented in the figure and, in fact, be bowl-shaped. In this figure, the horizontal axis is disjunction points and the vertical axis is cumulative percentage of samples, which are located under these disjunction points. On the other hand, the linear model closer to the best line, namely "BEST", is a better model. In training, the regressive decision tree was more effective (Figure 2). In addition, models' precision with environmental variables for training, test, and validation data are presented in Table 5.

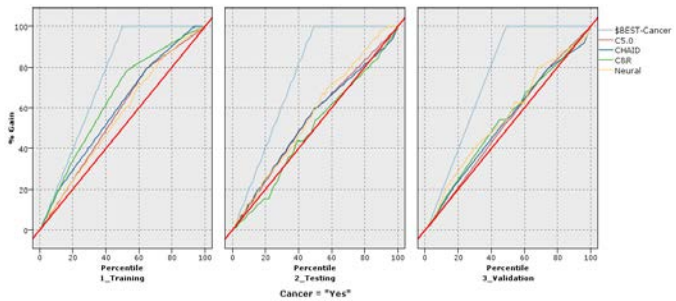

Figure 2: ROC figure of environmental variables

Table 5: Model's precision in training, test and validation (environmental variables)

Model's name	Model's precision in training %	Model's precision in test %	Precision of validation %
C5.0	65	60	55
CHAID	65	57	55
C & R	76	54	57
Neural Net	61	61	57

Discussion

In this research, the results were formed according to whether patients had cancer or not. In fact, cancer is the target variable in modeling stage. There were many variables for any cancerous patient, therefore understanding the effect of variables on each other helps us to describe the data better. In addition, in order to discover primary knowledge of these data, the statistical analyses and illustrations were performed on data, and by the help of these analyses, some rules were extracted in this research:

The extracted rules and results from clinical variables in CART, CHAID, C5.0 and Neural Net algorithms:

If a person does nott have pulmonary nodules and effusion of pleura fluid be equivalent to lack and/or closeness of pleural fluid, and equals one third of haemothorax; thus, with confidence of 100%, he/she is not affected by lung cancer (with 87 cases).

If a person has pulmonary nodules; thus, he/she is affected by lung cancer at confidence of 100% (with 64 cases).

If a person does not have pulmonary nodules and effusion of pleura fluid be equivalent to two thirds of haemothorax and all of haemothorax; thus, he/she is affected by lung cancer at confidence of 100% (with 23 cases). Meanwhile, there is no rule extracted for Neural Net algorithm.

In C5.0 and CHAID models, the important fields (risk factors) are pulmonary nodules and fluid of pleura effusion. Furthermore, pulmonary nodules and fluid of pleura effusion, hospital ward of patient admission, cough, diameter of pulmonary nodules, incidence of cancer among first degree relatives, history of non-pulmonary diseases especially cancer, and place of pulmonary nodules are the most important fields(risk factors) for C&R model. In addition, the important fields for the Neural Net model include pulmonary nodules and fluid of pleura effusion, hospital ward of patient admission, cough, diameter of pulmonary nodules, age, shortness of breath, chest graph, findings of spirometry, sex, hemoptysis, history of non-pulmonary diseases especially cancer, and place of pulmonary nodules.

The extracted rules and results from environmental variables

The extracted rules and results of fifth model of C5.0 algorithm:

If the geographical area served is Tehran, Sardasht-Faw, Boushehr, Zahedan, Isfahan, Zanjan, Qom, Ardebil, Abadan and Ahwaz; thus, with the confidence of 71% the person is not affected by lung cancer.

If the geographical area served is Bandar Abbas, Kermanshah, Khuzestan, Kurdistan, Sardasht, Dezfoul, Khoramshahr, Shalamcheh and Dehlavieh, Khark, Noshahr, Karaj, Gilan, Khoramabad, Arak, and Piranshahr; then, the person is affected by lung cancer at confidence of 61%. In this model, the important field (risk factor) is geographical area served.

The extracted rules and results of sixth model of CHAID algorithm:

If the geographical area served is Tehran, Ilam, Kurdistan and Sardasht-Faw; then, the person is not affected by lung cancer at confidence of 64%.

If the geographical area served is Boushehr, Zahedan, Isfahan, Zanjan, Qom, Ardebil, Abadan and Ahwaz; consequently, the person is not affected by lung cancer at confidence of 100%.

If the geographical area served is Bandar Abbas, Sardasht, Dezfoul, Khark, Noshahr, Karaj and Piranshahr; then, the person is affected by lung cancer at confidence of 89%.

If the geographical area served is Kermanshah, Khuzestan, Kurdistan, Khoramshahr, Shalamcheh and Dehlavieh, Gilan, Khoramabad and Arak; then, the person is affected by lung cancer at confidence of 55%. In this model, the important field (risk factor) is geographical area served.

The extracted rules and results of C&R algorithm:

If the geographical area served is Tehran, Ilam, Hamadan, Sardasht-Faw, Boushehr, Zahedan, Isfahan, Zanjan, Qom, Ardebil, Abadan and Ahwaz, and, the time of quitting smoking equals 1, 3, 5, 8, 9, 10, 15, 16, 20 and 40 years respectively; thus, the person isn't affected by lung cancer at confidence of 76%.

If the geographical area served is Bandar Abbas, Kermanshah, Khuzestan, Kurdistan, Sardasht, Dezfoul, Khoramshahr, Shalamcheh and Dehlavieh, Khark, Noshahr, Karaj, Gilan, Khoramabad, Arak, and Piranshahr, and the time of quitting smoking equals 2, 3, 5, 7, 8, 17, 18, 20 and 25 years and, smoking time equals 15, 25, 35 and 60 years; thus, the person is affected by lung cancer at confidence of 68%. In this model, the important fields (risk factors) are geographical area served, time of quitting smoking, occupational pollutants, smoking and addiction. Meanwhile, there is no rule extracted for Neural Net algorithm. In this model, the important fields are occupational pollutants, smoking, and addiction.

In previous studies[9,15-21], the authors have pointed out the importance of using explorative and future-sighted studies for exactly determining of disease stage and its influence on the patients' pre-awareness through decision-making tree and neural networks methods. Moreover, results of studies carried out by Najafi et al. [22], and Shaban et al. [23] were in consistent with one of our results that "the geographical area served" variable has an influence on patients' quality of life. Accordance with our results, the results of other researches[24,25] indicated the effect of age rise on getting lung cancer as well as its incidence among men more than women. Additionally, Van Klaveren et al[26]., similar to the results of this research, specified that the existence of pulmonary nodules makes the diagnosis and prediction of cancer faster and more precise. Of course, there is not an agreement between some results of this research and previous reports[27,28]. They studied the influence of smoking on lung cancer via linear models and linear logarithm.

Conclusions

According to models' precision, the results showed that in clinical variables, C5.0, CHAID, C & R models were stable and suitable for detection of lung cancer. In C5.0 and CHAID models, "pulmonary nodules" and "fluid of pleura effusion" variables have the greatest impact on detection of lung cancer. Furthermore, in C&R model, variables including pulmonary nodules and fluid of pleura effusion, hospital ward of patient admission, cough, diameter of pulmonary nodules, incidence of cancer among first degree relatives, history of non-pulmonary diseases especially cancer, and place of pulmonary nodules have the greatest impact on detection of lung cancer. For Neural Net model, pulmonary nodules and fluid of pleura effusion, hospital ward of patient admission, cough, diameter of pulmonary nodules, age, shortness of breath, chest graph, findings of spirometry, sex, hemoptysis, history of non-pulmonary diseases especially cancer and place of pulmonary nodules were very effective on detection of lung cancer. Furthermore, with regards to environmental variables, classification and regression tree (CART or C & R) models were stable and suitable for detection of lung cancer. In C5.0 and CHAID models, variable of geographical area served has the greatest impact on detection of lung cancer. In C&R model, variables including geographical area served, time of quitting smoking, occupational pollutants, smoking and addiction were very important for detection of lung cancer. In Neural Net model, variable of occupational pollutants, smoking and addiction have the supreme impact on detection of lung cancer.

In this research, the existence of pulmonary nodules, effusion of pleural fluid, diameter of pulmonary nodules, and place of lung nodules were identified as very important variables, which have the greatest impact on detection of lung cancer. On the other hand, variables such as geographical area served, occupational pollutants, time of quitting smoking, and smoking have been identified as relatively important variables that paying attention to these factors brings about a reduction in lung cancer cases. More importantly, factors such as repugnant conditions and different threating behaviors (for example smoking and addiction), existence of environments polluted with ether, sulfuric acid and nitric acid, electrical industries, chemical history with radioactive substances, fireproof cotton, ether halogen, ionization ray, insulating, brake pad and mustard gas can decrease consciousness level and disturb individuals' performance and cause loss of their power in a limited time that isn't acceptable for organizations. Therefore, providing a healthy environment with high quality and safety for the individuals is of more priority. In relation to clinical factors, some items with higher importance can be type of behavior with patient as well as prognosis of successfulness of surgical operations in order to prepare individuals to fast return to their workplaces, determining the chance of successfulness of therapeutic methods on serious diseases for cancerous patients, and diagnosis of diseases according to types of available information (medical images, characteristics of probable patient) before its occurrence in order to avoid from imposing catastrophic costs on organizations and families.

Acknowledgments

We gratefully thank all participants as well as staff of Baqiyatallah Hospital for their sincere participation in this study.

Conflict of interest statement

The authors declare that present study has no conflict of interest.

References

1. Kasper D. *Harrison's Manual of Medicine*. 18th ed. New York: Mc Graw-Hill; 2012.

2. Etemadi A, Semnani SH, Nouraie SM, Khademi H, Bahadori M. Cancer registry in Iran: a brief overview. *Arch Iranian Med.* 2008;11(5):577-580

3. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J Clin.* 2014;64(1):9-29.

4. Mousavi SM, Ramazani R, Davanlou M, Hajsadeghi N, Seddighi Z. Cancer incidence and mortality in Iran. *Ann Oncol.* 2009;20(3):556-563.

5. Bahader Y, Jazieh A. Epidemiology of Lung Cancer. *Ann Thorac Med.* 2008;2(3):65-67.

6. Danny R. Youlden PB. The international epidemiology of lung cancer, geographical distribution and secular trends. *J Thorac Oncol.* 2008;3(8):819-831.

7. Sadjadi A, Malekzadeh R, Derakhshan M H, Sepehr A, Nouraie M, Sotoudeh M, et al. Cancer occurrence in Ardabil: results of a population-based cancer registry from Iran. *Int J Cancer.* 2003;107(1):113-118.

8. Levin B, Lieberman D, McFarland B, Robert A. Smith D, Kimberly S, et al. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA Cancer J Clin.* 2008;58(3):130-160.

9. Khedmat H, Amini M, Izadi M, Nazareth MH, Qayyum H. The probability of survival of patients with gastric cancer and other official personnel of the armed forces hospital bagiyatallah in 1999 to 2004. *Mil Med.* 2006;9(3):167-177.

10. Krause N, Lynch J, Kaplan GA, Cohen RD, Goldberg DE, Salonen JT. Predictors of disability retirement. *Scand J Work Environ Health.* 1997;23(6):403-413.

11. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *Ai Mag.* 1996;17(3):37-54.

12. Chen H, Fuller SS, Friedman C, Hersh W. Knowledge management, data mining, and text mining in medical informatics. *Med Inform.*2005;8:3-33.

13. Collins GS, Mallett S**,** Omar O**,** Yu Ly-Mee. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* 2011;9:103.

14. Karimian Z, Kojuri J, Sagheb MM, Mahboudi A, Saber M, Amini M, et al. Comparison of residents' approaches to clinical decisions before and after the implementation of Evidence Based Medicine course. *J Adv Med Educ Prof.* 2014;2(4):170-175.

15. Esmaeilpour M, Naderifar V, Shukur Z. Cellular Learning Automata Approach for Data Classification. *Int J Innov Compt I.* 2012;8(12):8063-8076.

16. Khiabani I, Shahrabi J, Aliannejhad R. Finding sulfur mustard effect patterns on chemical injured people by data mining techniques. 3rd Data Mining Conference, Iran University of Science and Technology, Tehran; 2009.

17. Tavasoli S. Optimal Gene selection for cancer classification improvement. 4th Data Mining Conference, Sharif University of Technology, Tehran; 2009.

18. Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. Lung cancer survival prediction using ensemble data mining on SEER data. *Scientific Programming*. 2012;(20):29-42.

19. Delen D. Analysis of cancer data: a data mining approach. *Expert Syst.* 2009;26(1):100-112.

20. Murti S. Using Rule Based Classifiers for the Predictive Analysis of Breast Cancer Recurrence. *J Inform Engin Appl.* 2012;2(2):12-20.

21. Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. *Int J Com Appl.* 2012;47(10):44-48.

22. Najafi F, Jafari R, Mozafari HR, Leghaei Z. Trend In Lung Cancer Incidence In Kermanshah Province –Iran. *J Kermanshah Univ Med Sci*. 2010; 14(4): 342-348.[Persian]

23. Shaban M, Monjamed Z, Mehran A, Hasanpour Dehkordi A. The relation between the cancer characteristics and quality of life in the patients under chemotherapy. *Hayat*. 2004; 10 (3) :79-84.[Persian]

24. Montazeri B, Alizadeh K, Zarei S. The frequency of disabling malignancies in air force personnel during 1992-2001. *J IRIAF Health Admin.* 2007;10(2):5-11.

25. Almasi-Hashiani A, Farahmand M, Hassanzadeh J, Raei M. Lung cancer incidence: data from Fars Province cancer registry, 2001-2008. *Payesh.* 2012;11(4):477-483.[Persian]

26. Van Klaveren R, Oudkerk M, Prokop,M, Scholten T, Nackaerts K, Vernhout R, et al. Management of Lung Nodules Detected by Volume CT Scanning. *N Engl J Med*. 2009;361:2221-2229.

27. Liddell FA, Armstrong BG. The combination of effects on lung cancer of cigarette smoking and exposure in Quebec chrysotile miners and millers. *Ann Occup Hyg.* 2002;46(1):5-13.

28. Shouman M, Turner T, Stocker R. Applying k-nearest neighbour in diagnosing heart disease patients. *Int J Inf Edu Tech.* 2012;2(3):220-223.