RESEARCH ARTICLE

Analyzing a Lung Cancer Patient Dataset with the Focus on Predicting Survival Rate One Year after Thoracic Surgery

Peyman Rezaei Hachesu, Nazila Moftian, Mahsa Dehghani, Taha Samad Soltani*

Abstract

Background: Data mining, a new concept introduced in the mid-1990s, can help researchers to gain new, profound insights and facilitate access to unanticipated knowledge sources in biomedical datasets. Many issues in the medical field are concerned with the diagnosis of diseases based on tests conducted on individuals at risk. Early diagnosis and treatment can provide a better outcome regarding the survival of lung cancer patients. Researchers can use data mining techniques to create effective diagnostic models. The aim of this study was to evaluate patterns existing in risk factor data of for mortality one year after thoracic surgery for lung cancer. **Methods:** The dataset used in this study contained 470 records and 17 features. First, the most important variables involved in the incidence of lung cancer were extracted using knowledge discovery and datamining algorithms such as naive Bayes, maximum expectation and then, using a regression analysis algorithm, a questionnaire was developed to predict the risk of death one year after lung surgery. Outliers in the data were excluded and reported using the clustering algorithm. Finally, a calculator was designed to estimate the risk for one-year post-operative mortality based on a scorecard algorithm. **Results:** The results revealed the most important factor involved in increased mortality to be large tumor size. Roles for type II diabetes and preoperative dyspnea in lower survival were also identified. The greatest commonality in classification of patients was Forced expiratory volume in first second (FEV1), based on levels of which patients could be classified into different categories. **Conclusion:** Development of a questionnaire based on calculations to diagnose disease can be used to identify and fill knowledge gaps in clinical practice guidelines.

Keywords: Data mining- lung neoplasms- cancer survival- informatics- knowledge

Asian Pac J Cancer Prev, **18 (6)**, 1531-1536

Introduction

Data mining, as a new concept introduced in the mid-1990s, can help researchers gain new, profound insights and facilitate access to unanticipated knowledge sources in biomedical datasets (Yoo et al., 2012). Data mining is defined as a process to find patterns and connections in a database and to use data to build prediction models (Samad Soltani et al., 2015). Data mining is also regarded as a process for selecting, exploring and building models using the mass of data stored to discover prefabricated patterns (Fayyad et al., 1996). Data mining is used to identify new, accurate, understandable and potentially useful relationships and patterns within the data using a combination of data collection and extracting complex patterns to humans (Koh and Tan, 2011). It is very difficult to discover relationships between data sources in traditional statistical methods (Lee et al., 2000). Data mining has rapidly found many applications in wide areas such as healthcare delivery organizations, financial forecasting and weather forecasting (Obenshain, 2004) and most recently, in medical diagnostics (Einipour, 2011).

In healthcare, data mining is a very important branch in more profound diagnosis and perception of medical data and also, regarded an important field of research in prediction of diseases. Healthcare data mining intends to resolve real world healthcare issues in the diagnosis and treatment of diseases (Liao and Lee, 2002). Researchers use this method to diagnose various diseases and to do this, benefit from various methods and algorithms that differ in accuracy and precision (Shouman et al., 2012; Nabaei et al., 2016). Much of the issues in the medical field are concerned with the diagnosis of diseases based the tests conducted on the patient. Statistical studies mostly use data mining techniques to create classification models (Einipour, 2011; Samad-Soltani et al., 2015). Thus, this research was conducted to discover and report tacit knowledge in the selected cancer data, by analyzing the data collected for thoracic surgery using data mining and knowledge discovery methods.

Risk assessment for thoracic surgery in lung cancer and its associated data

According to official statistics published in 2013 in

*Department of Health Information technology, School of Health management and Informatics, Tabriz University of Medical Sciences, Tabriz, Iran. *For Correspondence: taha.soltany@gmail.com*

the United States, lung cancer had been the second most common cancer among both women and men (Siegel et al., 2013; Urvay et al., 2016), and the leading cause of death in North American countries including the United States and Canada (Jaklitsch et al., 2012). Globally, lung cancer is the most common cause of cancer-induced death in both men and women (Ridge et al., 2013; Satar et al., 2016). In Tehran, Iran, the age standardized incidence rate of lung cancer was 7.0 and 14.9 in women and men, respectively. it shows a relatively high incidence of lung cancer in comparison with other Asian countries. In addition, lung is a common site of mortal metastases (Ferlay and Whelan; Mohagheghi et al., 2009).

Although early diagnosis and treatment has a better outcome in the survival rate of cancer patients, however, issues arising from treatment methods can have negative effects on health-related quality of life (Gokgoz et al., 2011). One treatment method with side effects is thoracic surgery, which refers to any surgery on an organ and tissue within the chest including the lung. This surgery may be proposed when conventional treatments prove impractical or ineffective (Cerfolio et al., 2011). One of the main issues in clinical decision making for thoracic surgery is to select suitable patients for surgery by considering the risks and benefits of the action for them, both in short-term (postoperative) and longer term (postoperative surgery) (Zięba et al., 2014).

Many reports have pointed out the effects of gender and other demographic characteristics of individuals on lung cancer, and the survival rate in this type of cancer is highly affected by age, gender, tumor size, histology and tumor classification (Ferguson et al., 2000; Minami et al., 2000; Ludwig et al., 2005).

Materials and Methods

Method

This study was based on the data compiled from patients with lung cancer referring to health care centers that is also available in UCI datasets. This dataset had been collected retrospectively during the years 2007-2011 and registered in the Polish National Cancer Registry. The dataset contains 17 variables, as shown and described in Table 1(Zięba et al., 2014).

The selected dataset was a relative clean data which was some preprocess steps were performed by publishers. The dataset, included 470 records and consisted of 16 discrete input and one discrete output features. Each of the features listed in Table 1 has qualitative or quantitative range of values. The output classes also included two classes of 0 or 1, which respectively represented death or life.

Data mining stages were applied based on the following proposed model to discover patterns and investigate relationships in the dataset, on which the problem recognition, data collection and conversion and cleanup were performed by researchers in previous studies (Zięba et al., 2014; Zięba and Tomczak, 2015). This study, however, analyzed key influencers, detected categories and highlighted exceptions in the aforementioned dataset. Finally, a prediction calculator was proposed in form of a questionnaire to diagnose lung cancer based on the data. The tools used for analysis and data mining included the SQL server 2012 data mining add-ins for Excel. After installing the add-ins in the Microsoft Excel environment, two new tabs entitled Analyze and Data mining are added, and the options available in the analyze tab are in accordance with the categories of analyses performed in this study.

To analysis key influencers, which determines which factors had the strongest influence on the outcome columns, Microsoft naïve Bayes algorithm with automatic default parameters was used. Before than all continues features were discretized because Naive Bayes only accepts discrete attributes. To detect categories, Microsoft clustering algorithm was applied. the Expectation Maximization (EM) method was selected to clustering cases. This algorithm iteratively refines an initial cluster model to fit the data and determines the probability that a data point exists in a cluster (Jin and Han, 2011). Default heuristic parameters were selected to clustering. Exception highlighting was performed based on EM clustering algorithm and outliers were detected. The tool automatically sets this threshold for the initial analysis pass in detecting exceptions. To create scorecard calculator, Microsoft logistic regression algorithm was applied. This algorithm can work with discrete values, as well as discretized and continuous numeric data. Our dataset was discretized in preprocessing step, as well as missing data handling by average value imputation (Silva-Ramírez et al., 2011). To achieve more details about Microsoft algorithms, we googled "Microsoft Algorithm Technical Reference" (Guyer, 2016).

Results

Findings can be summarized in the two sections of data analysis and questionnaire creation. The results have been described below based on the time of the analysis performance:

Table 2 shows the results of the analysis of the key influencers. Key factors on the existing data on thoracic surgery are defined as follows:

"What are main key factors and symptoms of dead people after a year and living people in the dataset?". Table 2 shows the results of the key influencers. it composed of four columns. The first column contains variables existing in the dataset which represent identity and clinical data of patients referring to hospitals. The second column shows the values of the intended column. The third column is composed of zero or one that reflects the state of one's survival, with 1 demonstrating death after one year and zero demonstrating survival. The fourth column is the relative impact of the first column with the value of the second column for each row in the incidence of the third column. Relative impact indicates the strength of the association of this attribute with the outcome. The value of the column indicates the probability that the factor contributes to the outcome; therefore, the highest the value (between 0 and 100), the stronger the association.

By analyzing the research problem categories, individuals are classified into groups based on indicators

Table 1. Thoracic Surgery Data Set

Features (16 inputs and one outcome)
Feature 1: specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any(DGN 3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
Feature 2: FVC
Feature 3:FEV1
Feature 4: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
Feature 5: Pain before surgery (T,F)
Feature 6: Hemoptysis before surgery (T,F)
Feature 7: Dyspnea before surgery (T,F)
Feature 8: Cough before surgery (T,F)
Feature 9: Weakness before surgery (T,F)
Feature 10: T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest)
Feature 11: Type 2 Diabetes mellitus
Feature 12: MI up to 6 months (T,F)
Feature 14: Smoking (T,F)
Feature 15: Asthma (T,F)
Feature 16: Age at surgery (numeric)
Features 17: Risk1Y: (T (True or Died (N: 70 patients)), F (False or live(N: 400 patients))

derived from the features, and features of the individuals are similar within these categories or clusters. For each category, only two factors with the greatest relative impact are calculated, and the name of the category is selected as the name for the most effective agent. Table 3 shows the detected categories along with the number of records clustered within them as well as two main factors for classification along with their relative impact.

By detection analysis of exceptions, existing records in the database are identified along with unusual or possibly erroneous data. By determining a threshold of 75%, only 25% of exceptional data are recognized. Lower value of the threshold results in greater number of records to be identified as exceptions. Table 4 shows the results of detection of exceptions with the threshold level of 75%. In table 4, the first column presents the factor involved in making an outlier and the second column shows the frequency of the factor.

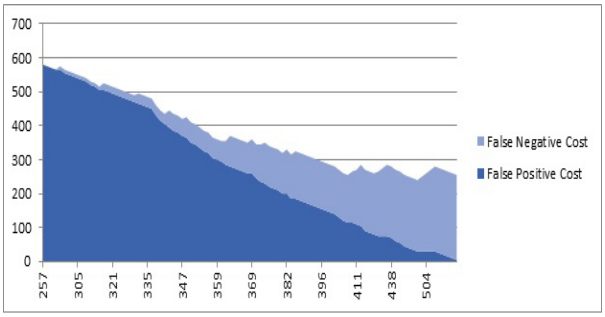

Figure 1. The Cost of Incidence of False Positives and False Negatives and the Threshold Value Based On Repetitive Sampling On the Dataset (Horizontal Axis Indicating the Calculated Threshold Value and Vertical Axis Indicating the Number of Samples); The Threshold

Table 2. Analysis Results of the Key Influencers on The Thoracic Surgery Dataset

Column	value	Status of Cancer	Relative impact
Type 2 Diabetes mellitus	False	False	36
Dyspnea before surgery	False	False	33
T in clinical TNM size of the original tumor	OC14	True	100
Type 2 Diabetes mellitus	True	True	36
Dyspnea before surgery	True	True	33

A questionnaire has been designed and developed for this dataset to identify people at risk of death by performing the prediction calculator analyses by means of data mining tools, which is presented in the results section. Due to the cost of the incorrect diagnosis of diseases (i.e., false negatives) and consequences risks for patients, the value of this variable was set as twice the value for detection of false positives which only included a further review cost (FN = 10, FP = 5). Figure 1 presents these values and the method to calculate the threshold value for separating the dead from the living people, which has been performed optimally and based on statistical methods. The threshold value used for distinguishing the costs was calculated as 482. Hence, high threshold values indicate greater postoperative mortality rate and lower threshold values indicate greater postoperative survival.

In Table 5, a questionnaire is proposed based on the Scorecard algorithm, which provides the ability to predict and detect without need for any software. In this table, an appropriate option is selected for the patient's condition based on the variables and their values which represent the intensity or scale of the variables of postoperative

Table 3. Analysis Results of Categories Detection for the Thoracic Surgery Dataset

Category name	No of records	Important factors	Factor value	Relative impact
Medium FEV1	194	FEV1	Medium	100
		FVC	High	20
Low FEV1 with high age	153	FEV1	Low	100
		Age	VeryHigh or, >69 years	39
Low FEV1 with low age	105	FEV1	Low	100
		Age	Low 48-55 years	19
High and Very high FEV1	18	FEV1	High	100
		FEV1	VeryHigh	100

Table 4. Results of Detection Analysis of Exceptions

Factor	Frequency
DGN	1
Hemoptysis before surgery	1
Total	2

Table 5. The Proposed Questionnaire To Determine The Risk Of Death According To The Variables Of The Studied Dataset. In this questionnaire, the first column presents the title of features, their ascending values and points of each option. If the sum of the values is higher than 482, the patient is at the risk of one-year postoperative mortality

DGN	DGN1	194
	DGN2	14
	DGN3	49
	DGN4	74
	DGN5	52
	DGN6	0
	DGN8	126
FVC	< 2.59	30
	2.59 - 3.18	69
	3.19 - 3.91	17
	3.92 - 4.66	33
	>= 4.67	0
FEV1	< 1.88	34
	1.88 - 2.24	31
	2.24 - 2.64	0
	2.64 - 3.24	52
	>= 3.24	5
Performance state	PRZ0	0
	PRZ1	8
	PRZ2	106
Pain before surgery	False	0
	True	33
Hemoptysis before surgery	False	0
	True	13
Dyspnea before surgery	False	0
	True	42
Cough before surgery	False	0
	True	3
Weakness before surgery	False	25
	True	0
T in clinical TNM	OC11	0
	OC12	28
	OC13	200
	OC14	206
type 2 Diabetes Mellitus	False	0
	True	25
MI up to 6 months	False	60
	True	0
peripheral arterial diseases	False	41
	True	0
Smoking	False	0
	True	21
Asthma	False	63
	True	0
Age	< 48	36
	48 - 55	12
	55 - 62	45
	62 - 69	0
	>= 69	9
		Total

mortality or survival. Then values of relative impact of each option are summed and if the final value exceeds the threshold value that is equal to 482, the patient has high risk of postoperative mortality and is very likely to die and vice versa. Numerical values of features are defined based on the scales defined in the dataset.

Discussion

The followings have been obtained based on the results of analyses performed on the dataset of thoracic surgery. With regard to the results obtained from the analysis of the key influencers, a direct relationship can be concluded between the non-incidence of type 2 diabetes in patients with lung cancer and their one-year postoperative survival, vice versa (meaning that, cancer patients with type 2 diabetes have high postoperative mortality). Moreover, patients without preoperative dyspnea have higher one-year postoperative survival, vice versa. However, based on the results of this analysis, the most important factor for increased operative mortality can be seen in the variables of TNM (Tumor, Node, and Metastasis) classification of malignant tumors, and tumor size. Accordingly, large tumors (OC14) lead to very high postoperative mortality. The results of this analysis can be used for case based reasoning or to develop clinical practice guidelines so that the existing knowledge gap in such guidelines while making medical decisions could be identified and resolved (Toussi et al., 2009). The results of analysis of the key influencers provide important features influencing the risk of death after thoracic surgery.

The analysis results of detecting the categories led to the classification of patients for future research. Accordingly, all the patients enrolled in the study were grouped based on their FEV1 value. However, this factor is not considered alone in the clustering and rather, is calculated along with other factors. These categorizations facilitate further processing on the data at a higher level. For future studies, such categorization is recommended to be applied on data with the features listed in the research. This categorization and aggregation of similar data in groups has been previously performed in studies (Louviere and Woodworth, 1983; Motschnig-Pitrik, 1996).

In analysis of detection of exceptions, two values are very rarely found in the dataset and form the minimum statistical sample. The first variable is DGN, which indicates that the record should be re-examined since it has a significant difference with other data and may be an outlier. However, the second variable is the history of hemoptysis for a patient prior to surgery that is

unusual and may also be an outlier. According to the two exceptions existing in the dataset, it can be argued that the data is considered to be low in error and high in quality.

Many studies have been conducted and evaluated to extract influential factors in determining surgical risk (Edwards et al., 1997; Nashef et al., 1999; Qaseem et al., 2006). The questionnaire, which was developed in this study based on a prediction calculator, should also be evaluated. However, the questionnaire is observed to certainly have the best performance on the existing datasets for the separation of the risk factors for one-year postoperative mortality and is regarded as a good measure. Therefore, the questionnaire can be used as a clinical diagnostic support tool after undergoing standardized evaluation and validation. In previously conducted studies, questionnaires have been designed and evaluated based on the Scorecard algorithm, which are often based on fewer and more general questions and therefore, their detection range is limited (Shin et al., 2010). In future studies, the questionnaire could be assessed and validated.

In conclusion, the method proposed along with the results in this study evaluates a framework to assess correlation and relationships between symptoms, risk factors and outcomes of diseases in relevant patients. Moreover, another practical application of this study could be the development of a questionnaire based on calculations to diagnose these diseases or their associated risks based on diagnostic costs, which could be subsequently used to identify and fill knowledge gaps in clinical practice guidelines. This method, although mostly used in the field of finance and sales, is functional in the area of healthcare, especially to examine individuals with diseases and thus, can be used to predict and classify patients.

Acknowledgements

Not applicable.

References

Cerfolio RJ, Bryant AS, Skylizard L, et al (2011). Initial consecutive experience of completely portal robotic pulmonary resection with 4 arms. J *Thorac Cardiovasc Surg*, **142**, 740-6.

Edwards FH, Grover FL, Shroyer ALW, et al (1997). The Society of Thoracic Surgeons national cardiac surgery database: current risk assessment. *Ann Thorac Surg*, **63**, 903-8.

Einipour A (2011). A fuzzy-ACO method for detect breast cancer. *Glob J Health Sci*, **3**, 195.

Fayyad UM, Piatetsky-Shapiro G, Smyth P, et al (1996). Advances in knowledge discovery and data mining, AAAI press Menlo Park, pp 1-36.

Ferguson MK, Wang J, Hoffman PC, et al (2000). Sex-associated differences in survival of patients undergoing resection for lung cancer. *Ann Thorac Surg*, **69**, 245-9.

Ferlay J, Whelan S (2002). Age-standardized and cumulative incidence rates (three-digit rubrics). *IARC Sci Publ*, **8**, 515-704.

Gokgoz S, Sadikoglu G, Paksoy E, et al (2011). Health related quality of life among breast cancer patients: a study from Turkey. *Glob J Health Sci*, **3**, 140.

Guyer C (2016). Data Mining (SSAS) [Online]. Microsoft documentation: Microsoft. Available: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-ssas 2017].

Jaklitsch MT, Jacobson FL, Austin JH, et al (2012). The American association for thoracic surgery guidelines for lung cancer screening using low-dose computed tomography scans for lung cancer survivors and other high-risk groups. J *Thorac Cardiovasc Surg*, **144**, 33-8.

Jin X, Han J (2011). Expectation maximization clustering. In 'Encyclopedia of Machine Learning', Eds Springer, pp 382-3.

Koh HC, Tan G (2011). Data mining applications in healthcare. *J Healthc Manag*, **19**, 65.

Lee I-N, Liao S-C, Embrechts M (2000). Data mining techniques applied to medical information. *Med Inform Internet Med*, **25**, 81-102.

Liao S-C, Lee I-N (2002). Appropriate medical data categorization for data mining classification techniques. *Med Inform Internet Med*, **27**, 59-67.

Louviere JJ, Woodworth G (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *J Mark Res*, **23**, 350-67.

Ludwig MS, Goodman M, Miller DL, et al (2005). Postoperative survival and the number of lymph nodes sampled during resection of node-negative non-small cell lung cancer. *CHEST J*, **128**, 1545-50.

Minami H, Yoshimura M, Miyamoto Y, et al (2000). Lung cancer in women: sex-associated differences in survival of patients undergoing resection for lung cancer. *CHEST J*, **118**, 1603-9.

Mohagheghi MA, Mosavi-Jarrahi A, Malekzadeh R, et al (2009). Cancer incidence in Tehran metropolis: the first report from the Tehran Population-based Cancer Registry, 1998-2001. *Arch Iran Med*, **12**, 15-23.

Motschnig-Pitrik R (1996). Analyzing the notions of attribute, aggregate, part and member in data/knowledge modeling. *J Syst Softw*, **33**, 113-22.

Nabaei A, Hamian M, Parsaei MR, et al (2016). Topologies and performance of intelligent algorithms: a comprehensive review. *J Artif Intell Rev*, **6**, 1-25.

Nashef SA, Roques F, Michel P, et al (1999). European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg*, **16**, 9-13.

Obenshain MK (2004). Application of data mining techniques to healthcare data. *Infect Control Hosp Epidemiol*, **25**, 690-5.

Qaseem A, Snow V, Fitterman N, et al (2006). Risk assessment for and strategies to reduce perioperative pulmonary complications for patients undergoing noncardiothoracic surgery: a guideline from the American College of Physicians. *Ann Thorac Surg*, **144**, 575-80.

Ridge CA, McErlean AM, Ginsberg MS (2013). Epidemiology of Lung Cancer. *Semin Intervent Radiol*, **30**, 93-8.

Samad-Soltani T, Ghanei M, Langarizadeh M (2015). Development of a fuzzy decision support system to determine the severity of obstructive pulmonary in chemical injured victims. *Acta Inform Med*, **23**, 138.

Samad Soltani T, Langarizadeh M, Zolnoori M (2015). Data mining and analysis: Reporting results for patients with asthma. *Payavard Salamat*, **9**, 224-34.

Satar R, Ali A, Abraha W, et al (2016). Estimating the economic burden of lung cancer in Iran. *Asian Pac J Cancer Prev*, **17**, 4729.

Shin B, Cole S, Park S-J, et al (2010). A new symptom-based questionnaire for predicting the presence of asthma. *J Investig Allergol Clin Immunol*, **20**, 27-34.

Shouman M, Turner T, Stocker R (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *Int J Inf Educ Technol*, **2**, 220.

Siegel R, Naishadham D, Jemal A (2013). Cancer statistics. *CA Cancer J Clin*, **63**, 11-30.

Silva-Ramírez E-L, Pino-Mejías R, López-Coello M, et al (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, **24**, 121-9.

Toussi M, Lamy J-B, Le Toumelin P, et al (2009). Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. *BMC Med Inform Decis Mak*, **9**, 1.

Urvay SE, Yucel B, Erdis E, et al (2016). Prognostic factors in stage III non-small-cell lung cancer patients. *Asian Pac J Cancer Prev*, **17**, 4693-7.

Yoo I, Alafaireet P, Marinov M, et al (2012). Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*, **36**, 2431-48.

Zięba M, Tomczak JM (2015). Boosted SVM with active learning strategy for imbalanced data. *Appl Soft Comput*, **19**, 3357-68.

Zięba M, Tomczak JM, Lubicz M, et al (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Appl Soft Comput*, **14**, 99-108.