Thoracic Cancer ISSN 1759-7706

ORIGINAL ARTICLE

Analysis of the factors influencing lung cancer hospitalization expenses using data mining

Tianzhi Yu[1,2], Zhen He[1], Qinghua Zhou[2], Jun Ma[2] & Lihui Wei[2]

1 College of Management & Economics, Tianjin University, Tianjin, China
2 Tianjin Medical University General Hospital, Tianjin, China

Keywords
Data mining; hospitalization expenses; influencing factors; lung cancer.

Correspondence
Zhen He, College of Management & Economics, Tianjin University, Tianjin 300072, China.
Tel: +86 22 87401783
Fax: +86 22 87401810
Email: zhhe@tju.edu.cn

Received: 21 April 2014;
Accepted: 10 June 2014.

doi: 10.1111/1759-7714.12147

Thoracic Cancer **6** (2015) 338–345

Abstract

Background: Hospitalization expenses for the therapy of lung cancer are not only a direct economic burden on patients, but also the focus of medical insurance departments. Therefore, the method for classifying and analyzing lung cancer hospitalization expenses so as to predict reasonable medical cost has become an issue of common interest for both hospitals and insurance institutions.

Methods: A C5.0 algorithm is adopted to analyze factors influencing hospitalization expenses of 731 lung cancer patients. A C5.0 algorithm is a data mining method used to classify calculation.

Results: Increasing the number of input variables leads to variation in the importance of different variables, but length of stay (LOS), major therapy, and medicine cost are the three variables of greater importance. They are important factors that affect the hospitalization cost of lung cancer patients. In all three calculations, the classification accuracy rate of training and testing partition sets reached 84% and above. The classification accuracy rate reached over 95% after addition of the cost variables.

Conclusion: The classification rules are proven to be in accordance with actual clinical practice. The model established by the research can also be applied to other diseases in the screening and analysis of disease hospitalization costs according to selected feature variables.

Introduction

Lung cancer mortality rate ranks first among all tumors.[1] Previous studies of lung cancer have focused primarily on genetic testing,[2] diagnosis,[3] therapy method,[4] therapy effectiveness,[5] survival estimates,[6] and health resource utilization,[7] while few studies have examined the influencing factors of lung cancer hospitalization expenses. Hospitalization expenses for lung cancer therapy are not only a direct economic burden on patients, but also the focus of medical insurance departments. Control of hospitalization expenses is both a measure to urge hospitals to improve the quality of medical services and an effective way to relieve the disease burden of patients under a competitive medical market environment.

Currently, studies of medical costs have mainly focused on cost-benefit analysis,[8] single disease payment, and diagnosis related group (DRG) studies.[9] Disease cost-benefit analysis probes into the relationship between the total treatment cost of the disease and the therapeutic effect for patients in order to select the most appropriate therapy method, reduce medical costs, and improve treatment. Especially in studies of chemotherapy, cost-benefit analysis serves to choose low-cost, high-efficacy chemotherapy drugs by studying the relationship between the cost of chemotherapy drugs and the survival time of patients. Disease cost-benefit analysis is the economic evaluation of clinical practice, which pursues maximization of effectiveness with the given cost, or minimization of cost on the basis of equal effectiveness. However, such analysis does not consider the factors affecting health care costs, and the influence of different factors.

Single disease payment study is defined as the scientific formulation of a fixed reimbursement criterion for each disease on the basis of a unified classification of disease diagnosis.[10] It standardizes the utilization of medical resources, but it is limited to diseases with definite diagnosis, single treatment, and few complications.

DRG studies sort patients into groups by taking into account factors such as age, gender, primary diagnosis, length of stay (LOS), surgery, and complications.[11] The health policy

maker then designates a cost range for each group, and the cost of each group is of significant difference. Doctors determine or predict hospital cost standards according to the cost range of different groups, in order to control health care cost. For cancer patients, diagnosis of tumor stages is an important factor for therapy choice. For example, patients with advanced cancer usually cannot undergo surgery. They rely on palliative care so the costs are relatively low. For all patient-DRGs, patients are first sorted into categories by major diagnosis, and then into categories by surgical and medical treatment. Finally, patients are classified into subgroups according to the severity of illness and the possibility of death. However, the established grouping methods are not designed specifically for different diseases and overlook the influence of other factors on medical expenses in the therapy process.

The three medical cost methods mentioned do not explain how to select an influence factor and calculate influence degree. In this paper, the data mining method and samples of lung cancer are taken into account for discovering medical cost influence factors and influencing degree according to disease features. With the development of information technology, data mining technology has been widely used in the medical field,[12] such as in disease diagnosis,[13] Chinese traditional medicine knowledge discovery,[14] health condition assessment,[15] treatment effect analysis,[16] prediction of possibility of death,[17] survival prediction,[18] clinical pathway,[19] medical quality improvement,[20] and medical insurance.[21] The main objective of the paper is to explore the factors influencing the hospitalization expenses of lung cancer patients and the role of different variables in the expenses using data mining. The paper then formulates classification rules for medical expenses so as to provide a theoretical basis for the control of hospitalization expenses and formulation of policies.

Methods

Data mining refers to the process of discovering information and knowledge, which is hidden and/or unknown, but potentially useful, from excessive, incomplete, fuzzy, random data.[22] Data mining technology rose in the late 1990s, with a distinctive interdisciplinary nature. Combined with computer technology, data mining reduces restrictions and binding of data analysis methods on data. It includes classification, association, clustering. Classification algorithms include: logistic regression, C5.0, CART, CHAID, QUEST, and ANN; association algorithms include: Apriori, Carma, and Sequence; and clustering algorithms include K-means, two-step clustering and Kohonen. Compared to other classification algorithms, the C5.0 algorithm is used to generate a multi-branch decision tree and rule sets. The input variables can be categorical or numeric and the output variable should be categorical. It can also deal well with missing values. This paper adopts a C5.0 algorithm and uses IBM SPSS MOLDER 14.2 software to analyze factors that affect hospitalization expenses of lung cancer patients.

The theoretical basis of C5.0 algorithm is information theory. It utilizes an information gain ratio as the criteria to determine the best grouping variables and split points. The algorithm takes the output variable as information U, which is emitted by the information source and the input variables as a series of information V, received by information sink. Before the decision tree is established, the output variable is completely random for the information sink. The average uncertainty is:

$$Ent(U) = -\sum_i P(u_i)\log_2 P(u_i)$$

In the decision tree generating process, when the information sink receives information, taking into account the input variables T1, conditional entropy of T1 is:

$$Ent(U|T1) = \sum_j P(t1_j)\left(-\sum_i P(u_i|t1_j)\log_2 P(u_i|t1_j)\right)$$

Information gain is:

$$Gains(U, T1) = Ent(U) - Ent(U|T1)$$

Greater information gain indicates greater ability to eliminate the average uncertainty from the source to the sink. If there are many classifications of input variables V, the information entropy will be too large. C5.0 takes the information gain ratio as the criteria for selection, which not only considers the degree of information gain, but also takes into account the cost paid to obtain information gain. The mathematical definition of information gain ratio is:

$$GainR(U, V) = Gains(U, V) \Big/ Ent(V)$$

Thereby, C5.0 selects the variables with maximum information gain ratio as grouping optimal variables and the split point. The C5.0 algorithm takes into account the processing of the missing values and considers samples with missing values as temporarily excluded samples. Moreover, weighting adjustment is also employed. So the problem of missing tumor node metastasis (TNM) stage values in this study is well solved.

Data

Medical records of lung cancer patients from a comprehensive hospital and a special cancer hospital of Tianjin during the period July 2012 to July 2013 were used as the data source of this paper. Through a review of inpatient medical record summaries, pathology reports, and discharge, admission, treatment process and surgery records, we obtained personal information, LOS, severity of illness, therapy choices, and hospitalization costs. Sample selection rules included: the first diagnosis must be lung cancer; an International

Classification of Diseases (ICD) code of D38.1; preoperative lung cancer; postoperative radiotherapy, chemotherapy, biological therapy medical records; and an additional ICD code of Z51.0, Z51.1, Z51.2, Z51.5, or Z51.8. Of 1327 samples, 731 samples were eligible for the study.

Feature indicators were extracted according to patient, diagnosis, main therapy, and severity of disease dimensions, respectively. Patient dimension included age and gender. Diagnosis dimension refers to the classification of lung cancer into small cell carcinoma, with a high degree of malignancy and poor prognosis; and non-small cell lung cancer, including squamous carcinoma, adenocarcinoma, adenosquamous carcinoma, large cell carcinoma, with a comparatively low degree of malignancy, according to the pathological classification.[23] Samples without detailed pathology reports were included in the "none of specialized" category (NOS). The main therapy dimensions included surgery, radiotherapy, chemotherapy, biological therapy, targeted therapy, and symptomatic treatment. While surgical treatment of lung cancer is the main therapy, the therapeutic effect of surgery is poor and it is highly risky if the lung cancer spreads to the mediastinum, heart, main bronchus, or has remote metastasis. Doctors, therefore, choose radiotherapy or chemotherapy as treatment. Radiotherapy and chemotherapy can also be adopted as preoperative therapies to reduce the risk of surgery. Targeted therapy drugs are often bought at outpatient departments and, therefore, were not included in this study. Symptomatic treatment deals with treatment of cough and lung infections of patients with lung cancer, or examination and treatment before the patients are diagnosed with lung cancer.

The severity of disease dimensions includes LOS, intensive care unit (ICU) time, number of secondary diagnoses, admission time, and TNM stage. LOS is an important indicator of health care resource utilization. The more complicated the disease, the longer patients stay in hospital and the more medical resources patients consume. The time that patients stay in the ICU is a direct indicator of severity of disease. Because the cost of the ICU is higher than the cost of a general ward, it leads to a difference in medical costs.

The number of secondary diagnoses reflects how many complications other than the major diagnosis are present. Some studies use the Charlson comorbidity index to measure the impact of the complications to the major diagnosis.[24] All-patient DRGs classified the severity of secondary diagnosis, which is not associated with the major diagnosis, into four levels. Though there is no direct correspondence between the severity of secondary diagnosis and all-patient DRGs, the severity of disease related groups should undergo a comprehensive assessment according to age, major surgery, or operation and major diagnosis. This paper calculates the number of secondary diagnoses, which is a direct indicator of other diseases of the patients. Admission time reflects how many times a patient has been hospitalized, which is particular to sustained cancer treatment. TNM stage is an important indicator of the severity of the tumor: "T" is based on the size of lung cancer, spread and location in the lung, and the extent of spreading to adjacent tissues; "N" indicates lymph node spread; and "M" refers to metastasis, spreading to distant organs. Once T, N, and M staging are clear, doctors can determine a clear comprehensive staging of 0, I_a, I_b, II_a, II_b, III_a, III_b or stage IV. Patients that belong to a relatively low stage have a good prospect for survival. Taking into consideration the number of samples of each TNM stage, the I_a, I_b stage are defined as stage I, II_a, II_b stage are defined as stage II, and III_a, III_b stage are defined as stage III.

The total hospitalization expense is taken as a decision variable, which includes medication, Chinese patent medicine, herbs, surgery, nursing, laboratory and examination fees and bed charges. Because the herb fee represents a low proportion of the total expense, herb fees are usually incorporated into the Chinese patent medicine fee, collectively referred to as the traditional Chinese medicine fee. Thus, seven types of fees constitute the total hospitalization expense. As a result of the implementation of a government-guided price, the hospital could not change drug, examination, laboratory or bed charges after approval by the price control bureau, therefore, the treatment cost of patients with lung cancer in the two hospitals in this article can be compared.

As shown in Figure 1, hospitalization expenses are analysed from five dimensions. There are also five categorical variables in Table 1 and twelve continuous variables in Table 2. The statistics data is calculated from 731 samples.

Results

Applying a C5.0 algorithm, according to mean-standard deviation grouping, this paper takes the total cost mean of the target variable plus or minus one, two, or three standard deviations as the group limit. The binning results were: (−∞ −30187.06), [−30187.06 −12749.85), [−12749.85 4687.36), [4687.36 39561.78), [39561.78 56998.99), [56998.99 744261.21), [744261.21 ∞). The sample is divided into training and testing sets with the proportion of 70% to 30%, respectively.

Step 1. Hospitalization expenses for patients with lung cancer were predicted. The training and testing set classification accuracies are 84.91% and 84.58%, respectively, solely according to patient, diagnosis, main therapy, and severity of disease dimensions. The top five influencing factors and their degree of importance are: LOS (0.68), main therapy (0.12), ICU time (0.07), admission time (0.06), and condition upon discharge (0.04). Kappa testing is used to analyze the classification outcome. The Kappa coefficient is 0.53, which means discrimination classification by data mining is of moderate coherence to the actual classification.

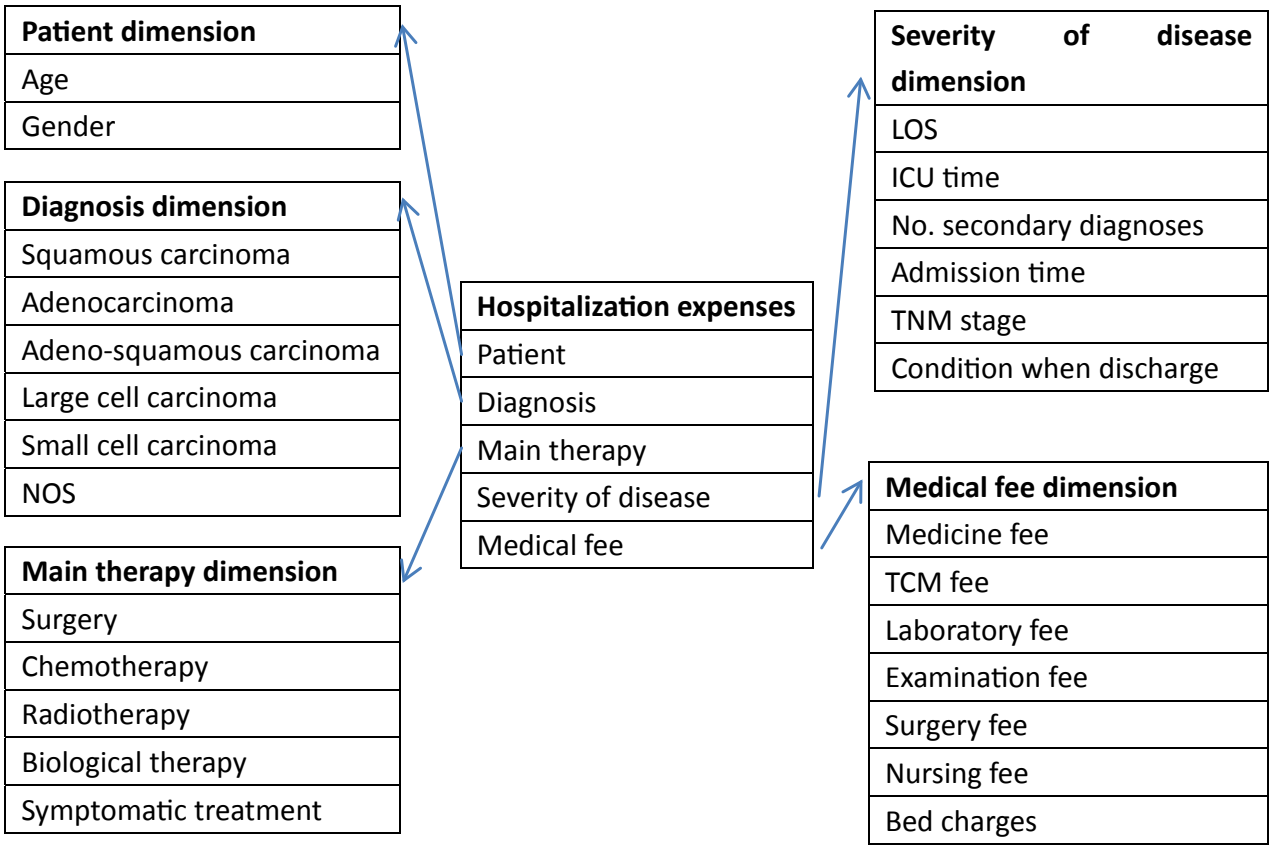

Figure 1 Variables star schema. LOS, length of stay; ICU, intensive care unit; TCM, traditional Chinese medicine; NOS, none of specialized category; TNM, tumor node metastasis.

Seventeen rules are concluded:

If LOS <= 7 and main therapy is chemotherapy, then the total expense belongs to [4687.36 39561.78);

If LOS <= 6 and main therapy is radiotherapy, then the total expense belongs to [−12749.85 4687.36);

If 6 < LOS <= 7 and main therapy is radiotherapy, then the total expense belongs to [4687.36 39561.78)

If LOS <= 7 and main therapy is biological therapy and TNM stage is 0 or 1, then the total expense belongs to [−12749.85 4687.36);

If LOS <= 7 and main therapy is biological therapy and TNM stage is 2, 3 or 4, then the total expense belongs to [39561.78 56998.99);

If 6 < LOS <= 7 and main therapy is symptomatic treatment, then the total expense belongs to [4687.36 39561.78);

If LOS <= 7 and admission time >7, then the total expense belongs to [−12749.85 4687.36);

If 7 < LOS <= 16 and admission time >7, then the total expense belongs to [4687.36 39561.78);

If 16 < LOS <= 30 and ICU time <= 158, then the total expense belongs to [4687.36 39561.78);

If 16 < LOS <= 30 and ICU time >= 158, then the total expense belongs to [39561.78 56998.99);

If LOS > 30 and main therapy is surgery, then the total expense belongs to [39561.78 56998.99);

If LOS > 30 and main therapy is surgery and discharge condition is death, then the total expense belongs to [56998.99 744261.21);

If LOS > 30 and main therapy is chemotherapy and TNM stage is 0 or 1 or 2 or 3, then the total expense belongs to [39561.78 56998.99);

If LOS > 30 and main therapy is chemotherapy and TNM stage is 4 and secondary diagnosis less than 2, then the total expense belongs to [39561.78 56998.99);

If LOS > 30 and main therapy is chemotherapy and TNM stage is 4 and secondary diagnosis more than 2, then the total expense belongs to [56998.99 744261.21);

If LOS > 30 and main therapy is radiotherapy or biological therapy, then the total expense belongs to [4687.36 39561.78);

If LOS > 30 and main therapy is symptomatic treatment, then the total expense belongs to [39561.78 56998.99).

Step 2. Only the medication fee input variable is added on the basis of Step 1, and the expenses of hospitalization

Table 1 Categorical variables

Variable	Description	No. (%)
Gender	Male	490 (67.0)
	Female	241 (33.0)
Diagnosis	Squamous carcinoma	188 (25.7)
	Adenocarcinoma	306 (41.9)
	Adeno-squamous carcinoma	26 (3.6)
	Large cell carcinoma	27 (3.7)
	Small cell carcinoma	123 (16.8)
	NOS	61 (8.3)
Main therapy	Surgery	180 (24.6)
	Chemotherapy	320 (43.8)
	Radiotherapy	70 (9.6)
	Biological therapy	98 (13.4)
	Symptomatic treatment	63 (8.6)
Condition when discharge	Recover	3 (0.4)
	Improving	519 (71.0)
	Not-healed	10 (1.4)
	Death	3 (0.4)
	Transferred to another hospital	183 (25.0)
	Other	13 (1.8)
TNM stage	0 stage:TisN0M0	18 (2.8)
	I stage:T1N0M0, T2aN0M0	69 (10.5)
	II stage:T2bN0M0, T1N1M0, T2aN1M0, T2bN1M0, T3N0M0	172 (26.2)
	III stage:T1-2N2M0, T3N1-2 M0, T4N0-1 M0, T4N2M0, TanyN3M0	183 (27.9)
	IV stage:T any N any M1	214 (32.6)

NOS, none of specialized category; TNM, tumor node metastasis.

for patients of lung cancer are given reasonable judgment. The training set classification accuracy rate is 96.71%, and the testing set classification accuracy is 95.33%. The top five influencing factors and the degree of importance are: medication fee (0.86), LOS (0.09), main therapy (0.03), admission time (0.01), and age (0.00). The Kappa coefficient is 0.90, which means discrimination classification by data mining is of perfect coherence to the actual classification.

Twenty rules are concluded:

If medicine fee <= 1827.3 and LOS <= 8, then the total expense belongs to [−12749.85 4687.36);

If medicine fee <= 144.5 and LOS > 8, then the total expense belongs to [−12749.85 4687.36);

If 144.5 < medicine fee <= 1827.3 and LOS >8, then the total expense belongs to [4687.36 39561.78);

If 1827.3 < medicine fee <= 22969.3 and LOS <= 8, then the total expense belongs to [4687.36 39561.78);

If 1827.3 < medicine fee <= 22969.3 and LOS <= 5, then the total expense belongs to [−12749.85 4687.36);

If 1827.3 < medicine fee <= 22969.3 and 5 < LOS <= 8, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS <= 20, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 25681.2 and LOS > 20 and main therapy is surgery, and age <=63, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 25681.2 and LOS > 20 and main therapy is surgery, and age >= 63, then the total expense belongs to [39561.78 56998.99);

If 25681.2 < medicine fee <= 35373.7 and LOS > 20, and main therapy is surgery, then the total expense belongs to [39561.78 56998.99);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is chemotherapy and gender is male, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is chemotherapy and gender is female, then the total expense belongs to [39561.78 56998.99);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is radiotherapy, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is biological therapy, then the total expense belongs to [39561.78 56998.99);

Table 2 Continuous variables

Variable	Unit	Min	Max	Mean	Standard deviation
Age	Year	16	85	59	10
LOS	Day	1	122	16	13.1
ICU time	Hour	0	304	26.5	51.9
Other diagnosis	Number	0	7	1	1
Admission time	Number	1	34	3	3.4
Medication fee	CNY	0	124457	15738	13311.5
TCM fee	CNY	0	5257	590	650.5
Laboratory fee	CNY	0	20141	1991	2177.8
Examination fee	CNY	0	10000	1093	1540.3
Surgery fee	CNY	0	5750	648	1136.4
Nursing fee	CNY	0	513	52	63
Bed charges	CNY	75	24000	2013	2154.7

CNY, China Yuan; ICU, intensive care unit; LOS, length of stay; TCM, traditional Chinese medicine.

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is symptomatic treatment, then the total expense belongs to [39561.78 56998.99);

If 35373.7 < medicine fee <= 48559.3 and main therapy is surgery and age <=58, then the total expense belongs to [39561.78 56998.99);

If 35373.7 < medicine fee <= 48559.3 and main therapy is surgery and age >= 58, then the total expense belongs to [56998.99 744261.21);

If 35373.7 < medicine fee <= 48559.3 and main therapy is not surgery, then the total expense belongs to [39561.78 56998.99);

If 48559.3 < medicine fee <= 52661 and all, then the total expense belongs to [56998.99 744261.21);

If medicine fee > 52661 and all, then the total expense belongs to [744261.21 ∞).

Step 3. All cost dimension variables are included on the basis of Step 1. The training and testing set classification accuracies are 99.03% and 98.13%, respectively. The top five influencing factors and the degree of importance are: medication fee (0.78), LOS (0.04), nursing fee (0.04), main therapy (0.04), and examination fee (0.03). The Kappa coefficient is 0.96, which means discrimination classification by data mining is of perfect coherence to the actual classification.

Eighteen rules are concluded:

If medicine fee <1827.3 and examination fee <=1225 and nursing fee <=27, then the total expense belongs to [−12749.85 4687.36);

If medicine fee < 1827.3 and examination fee <=1225 and nursing fee > 27, then the total expense belongs to [4687.36 39561.78);

If 1827.3 < medicine fee <= 5719 and laboratory fee <=105, then the total expense belongs to [−12749.85 4687.36);

If 1876 <= medicine fee <= 5719 and laboratory fee > 105, then the total expense belongs to [4687.36 39561.78);

If 5719 <= medicine fee <= 22969.3 and laboratory fee <=105, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS <= 20, and bed charges <=4650, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS <= 20, and bed charges > 4650, then the total expense belongs to [39561.78 56998.99);

If 22969.3 < medicine fee <= 25681.2 and LOS > 20 and main therapy is surgery and bed charges <=3480, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 25681.2, LOS > 20 and main therapy is surgery and bed charges > 3480, then the total expense belongs to [39561.78 56998.99);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is chemotherapy and gender is male, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is chemotherapy and gender is female, then the total expense belongs to [39561.78 56998.99);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is radiotherapy, and laboratory fee <=4685, then the total expense belongs to [4687.36 39561.78);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is radiotherapy and laboratory fee > 4685, then the total expense belongs to [39561.78 56998.99);

If 22969.3 < medicine fee <= 35373.7 and LOS > 20 and main therapy is biological therapy or symptomatic treatment, then the total expense belongs to [39561.78 56998.99);

If 35373.7 < medicine fee <= 48559.3 and bed charges <=5564 and surgery fee <=4350, then the total expense belongs to [39561.78 56998.99);

If 35373.7 < medicine fee <= 48559.3 and bed charges <=5564 and surgery fee > 4350, then the total expense belongs to [56998.99 744261.21);

If 35373.7 < medicine fee <= 48559.3 and bed charges > 5564 and surgery fee <=2687, then the total expense belongs to [39561.78 56998.99);

If 35373.7 < medicine fee <= 48559.3 and bed charges > 5564 and surgery fee > 2687, then the total expense belongs to [56998.99 744261.21).

Discussions

We used a combination of different input variables and found that all classification accuracy rates reached 84% and above. The two variables of LOS and main therapy are always among the top five influencing factors. When the variable of medication fee is included in the study, it is the most important factor, which is in line with medical practice and demonstrates the effectiveness of the C5.0 algorithm model.

LOS is an important factor influencing the medical expenses of patients with lung cancer. As LOS is the best variable for grouping, classification of the hospitalization costs of patients with lung cancer can be more effective if LOS is combined with other variables. Shortening the LOS is an effective way to reduce the cost of treatment for patients with lung cancer. Hospitals need to use the latest medical technology, make a definite diagnosis of the disease as soon as possible, and implement clinical pathway management to limit LOS. Medical insurance departments may try to implement the same reimbursement rate for outpatient and hospitalization costs and encourage patients to take examinations, such as medical imaging and bronchoscopy, in outpatient departments before they are hospitalized, in order to avoid factors such as examination and laboratory tests prolonging LOS. The different therapy types available to lung cancer patients have an obvious impact on hospitalization costs. For example, patients who need to undergo surgery

have to pay a surgery fee, and, in addition, intensive care service after surgery should also be paid. Before a patient's hospitalization, classification rules should be formulated following the first step we have outlined above. This can provide a reference for the choice of therapy. It is then possible for medical insurance departments to determine the appropriate range of hospitalization expenses for lung cancer patients through the classification rules and in order to prevent medical insurance fraud.

After inclusion of the variable of costs, medication fees become a major influencing factor of hospitalization expenses for patients with lung cancer, with the highest weight value in all of the variables. Medical insurance departments can verify if medical expenses are reasonable through the second and third steps. Medication has been the focus of attention for the patient, and the state health department also treats the lowering of drug prices as a key reform. Solving the problem of drug costs on the one hand depends on reducing intermediate links in circulation and the corresponding costs in the drug distribution system; on the other hand, it depends upon getting free medication through public non-profit organizations and reducing hospital expense on drugs. For lung cancer patients who have undergone surgery, or patients with advanced lung cancer, their intensive care stay is long, resulting in nursing fees and ICU time being influential factors on hospitalization costs; patients who have received ordinary chemotherapy, radiation therapy, biological therapy and symptomatic treatment do not incur these costs. Examination fees are also listed in the top five influencing factors, because lung cancer diagnosis and monitoring of progression relies mainly on computed tomography (CT), emission (E)CT, positron emission tomography-CT, and other expensive examination methods. Surgery fees have less impact on hospitalization costs, mainly because the price of surgery is low compared with drug and examination costs. This is not in line with the fact that lung surgery is of high risk, difficult, and requires advanced technical skill. Therefore, it is recommended that there should be an appropriate increase in surgery price to reflect the technical value of doctors.

Indicators of patient and diagnosis dimensions, such as gender, age, different pathological types of lung cancer diagnosis, or different TNM stage are unimportant variables and have a low impact on hospitalization costs.

This study has some limitations because variables, such as ethnic groups, residence, occupation, household income, and other demographic indicators are not included in the research because of a lack of demographic data in the patients' medical records. The impact of these variables on the hospitalization expenses of patients could not be confirmed and depends on the future improvement of health information systems to promote the comprehensiveness of data collection for further analysis.

Conclusions

The application of a C5.0 algorithm to determine the influencing factors and classification rules of the costs of treatment for lung cancer patients addresses the problems of predicting the expense of hospitalization and verification of the reasonableness of hospitalization expenses. It also a reference for patients in their choice of therapy and provides decision support for hospitals to reduce the cost of medical services. Moreover, it has provided a reference to medical insurance departments to determine the range of hospitalization expenses of lung cancer patients. The proposed method that this paper established takes into account the effects of different variables on the cost of inpatient treatment of lung cancer and the classification rules conform to clinical practice of lung cancer. The classification rules that the model generated reveal the hospitalization expenses of lung cancer patients and can be applied to other diseases. Compared with the established classification method, this model can screen different feature variables and utilize feature variables in hospitalization expenses analysis of various diseases according to their features.

Acknowledgment

We would like to thank Tianjin University and Tianjin Medical University for funding this project, 2011KY10. The paper is also financially supported by Natural Science Foundation of China (NSFC, 71225006).

Disclosure

No authors report any conflict of interest.

References

1 Plunkett TA, Chrystal KF, Harper PG. Quality of life and the treatment of advanced lung cancer. *Clin Lung Cancer* 2003; **5**: 28–32.
2 Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. *Comput Biol Med* 2007; **37**: 251–61.
3 Qiang Y, Guo Y, Li X, Wang Q, Chen H, Cuic. The diagnostic rules of peripheral lung cancer preliminary study based on data mining technique. *J Nanjing Med Univ* 2007; **21**: 190–5.
4 Eccles BK, Geldart TR, Laurence VM, Bradley KL, Lwin MT. Experience of first-and subsequent-line systemic therapy in the treatment of non-small cell lung cancer. *Ther Adv Med Oncology* 2011; **3**: 163–70.
5 Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol* 2013; **108**: 174.
6 Hendryx M, O'Donnell K, Horn K. Lung cancer mortality is elevated in coal-mining areas of Appalachia. *Lung Cancer* 2008; **62**: 1–7.

7 Phillips-Wren G, Sharkey P, Morss Dy S. Mining lung cancer patient data to assess healthcare resource utilization. *Expert Syst Appl* 2008; **35**: 1611–9.
8 Kim J, Lee E, Lee T, Sohn A. Economic burden of acute coronary syndrome in South Korea: a national survey. *BMC Cardiovasc Disord* 2013; **13**: 55.
9 Scheller-Kreinsen D, Quentin W, Busse R. DRG-based hospital payment systems and technological innovation in 12 European countries. *Value Health* 2011; **14**: 1166–72.
10 Ming C, Yan G. [Impact of single disease payment system on hospital delivery service providers' behavior.] *Beijing Da Xue Xue Bao* 2012; **44**: 387–91. (In Chinese.)
11 Averill RF, Goldfield N, Hughes JS *et al*. All patient refined diagnosis related groups: Methodology Overview (Version 20.0). 3M Health Information Systems, 2003.
12 Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inf* 2008; **77**: 81–97.
13 Alizadehsani R, Habibi J, Hosseini MJ *et al*. A data mining approach for diagnosis of coronary artery disease. *Comput Methods Programs Biomed* 2013; **111**: 52–61.
14 Zhou X, Chen S, Liu B, Zhang R *et al*. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artif Intell Med* 2010; **48**: 139–52.
15 Silva A, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. *Artif Intell Med* 2008; **43**: 179–93.
16 Jonsdottir T, Hvannberg ET, Sigurdsson H, Sigurdsson S. The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. *Expert Syst Appl* 2008; **34**: 108–18.
17 Silva A, Cortez P, Santos MF, Gomes L, Neves J. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artif Intell Med* 2006; **36**: 223–34.
18 Kusiak A, Dixon B, Shah S. Predicting survival time for kidney dialysis patients: a data mining approach. *Comput Biol Med* 2005; **35**: 311–27.
19 Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. *Artif Intell Med* 2012; **56**: 35–50.
20 Chae YM, Kim HS, Tark KC, Park HJ, Ho SH. Analysis of healthcare quality indicator using data mining and decision support system. *Expert Syst Appl* 2003; **24**: 167–72.
21 Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. *Int J Med Inform* 2001; **62**: 103–11.
22 Han J, Kamber M. *Data Mining Concepts and Techniques, 2nd Edition*. Morgan Kauffman, San Francisco 2006; 1–45.
23 Hosgood HD III, Farah C, Black CC, Schwenn M, Hock JM. Spatial and temporal distributions of lung cancer histopathology in the state of Maine. *Lung Cancer* 2013; **82**: 55–62.
24 Needham DM, Scales DC, Laupaeis A, Pronovost PJ. A systematic review of the Charlson comorbidity index using Canadian administrative databases: a perspective on risk adjustment in critical care research. *J Crit Care* 2005; **20**: 12–9.

