P05-2008

Data Mining and Analysis to Lung Disease Data

Guoxin Tang,University of Louisville,Louisville, KY

ABSTRACT

Objective: To examine the relationship between patient outcomes and conditions of the patients undergoing different treatments for lung disease.

Method: SAS Enterprise Guide was used to obtain Lung disease data from the NIS (National Inpatient Sample) by using CATX and RXMATCH statements. We first concatenate all 15 columns of diagnosis codes into one text string using the CATX statement. The RXMATCH looks for the initial code of '162' to find all patients with a diagnosis of lung disease. The ICD 9 code of 162 means malignant neoplasm of trachea, bronchus, and lung. Kernel Density Estimation was used to examine the lung disease by Age, Length of Stay and Total Charges, which showed the relationships among these outcomes by using data visualization. Then we use SAS Text Miner to investigate relationships in co-morbid diagnoses. To investigate, we used SAS Text Miner to define clusters of diagnoses. Then we can inspect the results by defining a severity measure using text analysis.

Results: After filtering by lung disease, there were more than 8000 observations in the data. The examination reveals that there was certainly a relationship between lung disease and Age, Length of Stay and Total Charges. Patients with lung diseases increase inpatient events starting at age 38, accelerating at age 45, and decreasing at 78. They have a higher probability of a stay of 4 days, which indicates that there was a higher probability of higher cost.

Conclusion: By using the Kernel Density Estimation and Text Miner, we obtained the statistical information about the Age, Length of Stay and Total Charges for patients with lung diseases. Cluster analysis also gave us five diagnosis clusters with ranking of severity by severity measure.

INTRODUCTION

The Nationwide Inpatient Sample (NIS) [1] is part of the Healthcare Cost and Utilization Project (HCUP), sponsored by the Agency for Healthcare Research and Quality (AHRQ), formerly the Agency for Health Care Policy and Research. The NIS is the largest all-payer inpatient care database that is publicly available in the United States, containing data from 5 to 8 million hospital stays from about 1,000 hospitals sampled to approximate a 20-percent stratified sample of U.S. community hospitals.

It is the only national hospital database with charge information on all patients, regardless of payer, including persons covered by Medicare, Medicaid, private insurance, and the uninsured. The NIS's large sample size enables analyses of rare conditions, such as congenital anomalies; uncommon treatments such as organ transplantation; and special patient populations such as the uninsured. It is the purpose of our study to examine the relationship between patient outcomes and investigate relationships in co-morbid diagnoses.

METHODS

The lung cancer data are from the NIS, and we had five years of data, 2000 to 2004. Each year of data included about 8 million records. Here, we first consider the data for the year, 2004.

In order to simplify the process of discovery, we first concatenate all 15 columns into one text string using the CATX statement. To find those patients with Lung Cancer, the RXMATCH function was used. This code put all possible diagnosis codes into one text string, and defined a second string containing all possible procedure codes using the CATX statement. The RXMATCH looked for the initial code of '162' that found all patients with a diagnosis code related to lung disease. Because '162' can occur in other codes that are not related to lung cancer, such as '216.2', we use four digits of code rather than three to avoid catching '216.2'. The code used was the following:

```
data lc.lungcancernew;
set lc. nis_2004_10pct_sample_a;
lungcancer=0;
diagnoses=catx(' ',dx1, dx2, dx3, dx4, dx5, dx6, dx7, dx8, dx9, dx10, dx11,
dx12, dx13, dx14, dx15);
procedures=catx(' ',pr1, pr2, pr3, pr4, pr5, pr6, pr7, pr8, pr9, pr10, pr11,
pr12, pr13, pr14, pr15) ;
if (rxmatch('1620',diagnoses)>0) then lungcancer=1;
if (rxmatch('1621',diagnoses)>0) then lungcancer=1;
if (rxmatch('1622',diagnoses)>0) then lungcancer=1;
if (rxmatch('1622',diagnoses)>0) then lungcancer=1;
if (rxmatch('1623',diagnoses)>0) then lungcancer=1;
if (rxmatch('1624',diagnoses)>0) then lungcancer=1;
if (rxmatch('1625',diagnoses)>0) then lungcancer=1;
if (rxmatch('1626',diagnoses)>0) then lungcancer=1;
if (rxmatch('1627',diagnoses)>0) then lungcancer=1;
if (rxmatch('1628',diagnoses)>0) then lungcancer=1;
if (rxmatch('1629',diagnoses)>0) then lungcancer=1;
run;
```

The ICD 9 code of 162 means malignant neoplasm of trachea, bronchus, and lung. Below are all of the 4-digit codes associated with lung diseases:

162.0 Trachea (Cartilage of trachea, Mucosa of trachea).
162.2 Main bronchus (Carina, Hilus of lung).
162.3 Upper lobe, bronchus or lung.
162.4 Middle lobe, bronchus or lung.
162.5 Lower lobe, bronchus or lung.
162.8 Other parts of bronchus or lung (Malignant neoplasm of contiguous or overlapping sites of bronchus or lung whose point of origin cannot be determined).
162.9 Bronchus and lung, unspecified.

There are a total of 8,216 observations for 2004 related to lung disease out of 800,000 records. Note that approximately 1.05% of the inpatient population has a diagnosis of lung disease (shown in Table 1).

Table 1. The frequency of lung diseases of 2004.

The FREQ Procedure

lungcancer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	792241	98.97	792241	98.97
1	8216	1.03	800457	100.00

Then Kernel Density Estimation (KDE procedure) was used to examine the lung disease by Age, Length of Stay and Total Charges, which showed the relationships among these outcomes by using data visualization.

In order to perform text analysis on the lung cancer data, the Text Miner node in Enterprise Miner was used to examine the data according to text strings of patient conditions. To define text clusters, we limit the number of terms to ten to describe clusters. We use the standard defaults of Expectation Maximization and Singular Value Decomposition. In order to compare outcomes by text clusters, we merge the cluster descriptions and the cluster numbers into the original dataset. We use kernel density estimation to make a comparison of age, length of stay and cost by clusters. The code was the following:

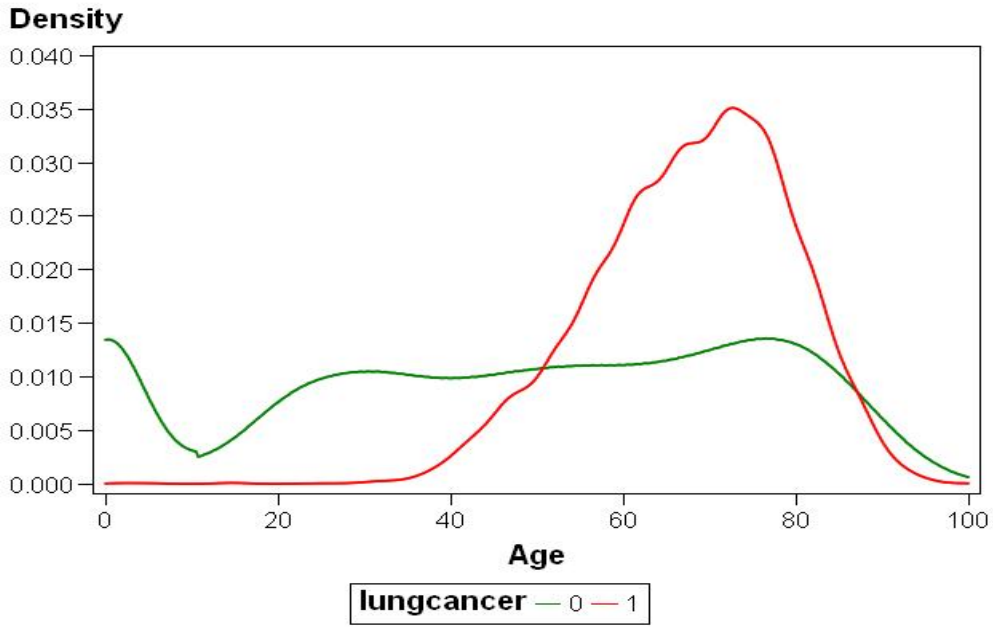
```
data emws1.clusternis (keep=_cluster_ _freq_ _rmsstd_
clus_desc);
set emws1.text_cluster;
run;
data emws1.desccopynis (drop=_svd_1-_svd_500
_roll_1-_roll_1000 prob1-prob500);
set emws1.text_documents;
run;
proc sort data=emws1.clusternis;
by _cluster_;
proc sort data=emws1.desccopynis;
by _cluster_;
data emws1.nistextranks;
merge emws1.clusternis emws1.desccopynis;
by _CLUSTER_;
run;
proc kde data=emws1.nistextranks;
univar totchg/gridl=0 gridu=100000
out=emws1.kdecostbycluster;
by _cluster_;
```

```
run;
proc kde data=emws1.nistextranks;
univar age/gridl=0 gridu=100
out=emws1.kdeagebycluster;
by _cluster_;
run;
proc kde data=emws1.nistextranks;
univar los/gridl=0 gridu=35
out=emws1.kdelosbycluster;
by _cluster_;
run;
```

RESULTS

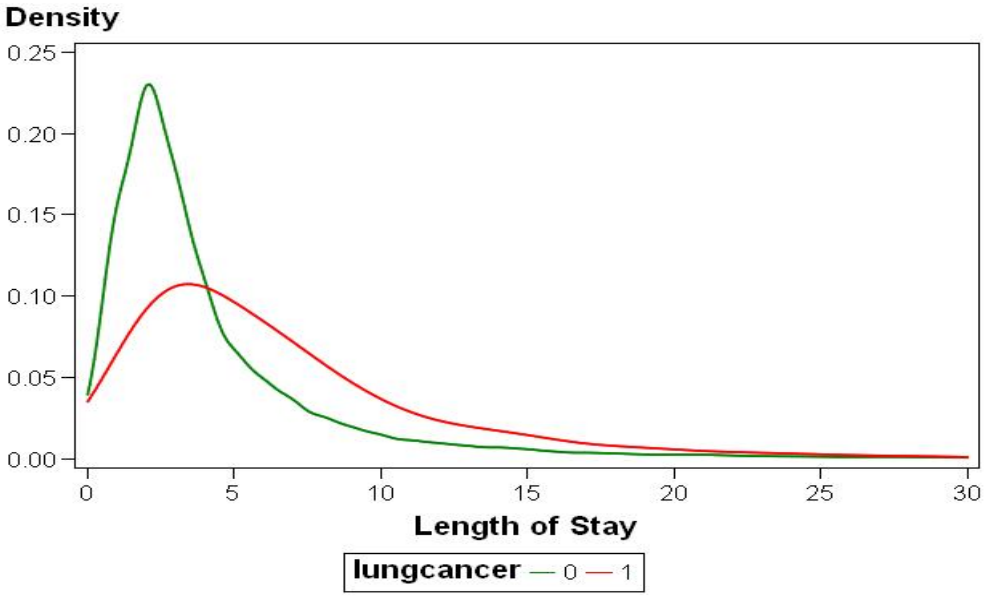
First, we use PROC KDE to examine the variables in relationship to the data with kernel density. The main advantage of using kernel density estimation is that the graphs can be overlaid for more direct comparisons. For example, we consider the relationship of lung cancer to Age, Length of Stay and Costs.

Figure 2. The Kernel Density of Lung cancer by Age using Kernel Density Estimation.



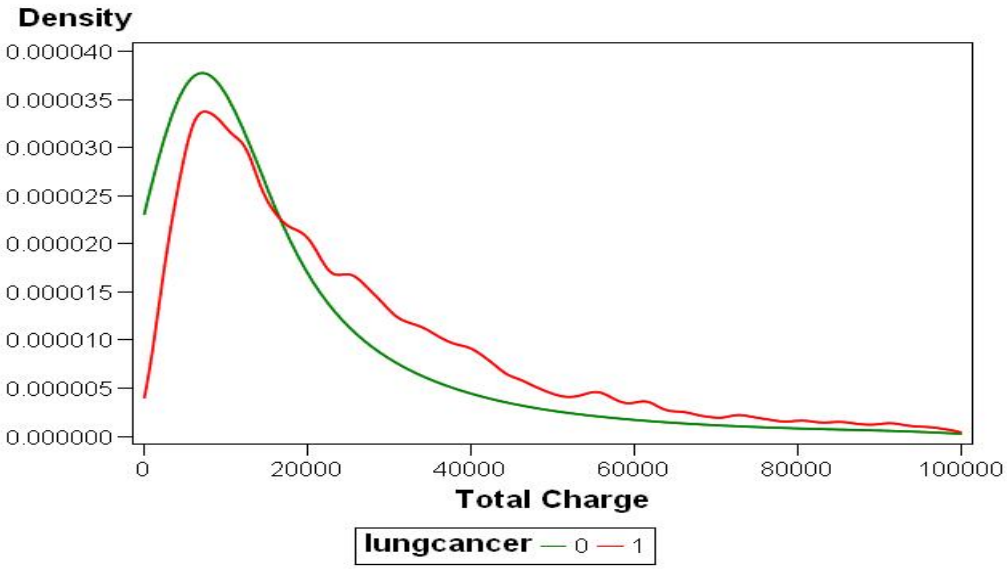
Note that the patients without lung diseases have a relatively constant likelihood of an inpatient event regardless of age (except for the interval of 0 to 20 and 60 to 80, where there is a slight change. However, patients with lung diseases increase inpatient events starting at age 38, accelerating at age 45, and decreasing at age 78.

Figure 3. The Density of Lung cancer by LOS (Length of Stay) using Kernel Density Estimation.



Those with lung diseases have a higher probability of a stay of 6 or more days, and a lower probability of staying 5 or fewer days compared to patients without lung diseases.

Figure 4: The density of Lung cancer by Total Charge using Kernel Density Estimation.



Note that there is an intersection point of costs for patients at around 19,000, indicating that there is a higher probability of higher cost if the patient has lung diseases.

Next, we considered the diagnosis codes and examined more in-depth the types of complications that patients have in relation to lung cancer. Recall that there are 8,216 patients with a diagnosis of lung cancer. Patients can be represented in multiple categories. Here, Text Miner in Enterprise Miner was used to examine the data according to text strings of patient conditions. Cluster analysis was used to find the categories of documents. For example, the text analysis defined seven different clusters in the data that were given in Table 2.

Table 2. Cluster table for diagnosis strings.

# ▲	Descriptive Terms	Freq	Percentage	RMS Std.
1	5990, 486, 2859, 25000, 42731	1236	0.15043816...	0.1259563...
2	25000, 41401, 4280, 412, 41400	938	0.11416747...	0.1186902...
3	1629, 486, 4280, 42731, 2765	1666	0.20277507...	0.1283921...
4	3051, 1623, 5121, 496, v1582	399	0.04856377...	0.1100431...
5	311, 1972, 1622, 53081, 3051	1387	0.16881694...	0.1285800...
6	1985, 2768, 1628, 2765, 1983	1641	0.19973222...	0.1242998...
7	3051, v1582, 1961, 1625, 49121	949	0.11550632...	0.1240970...

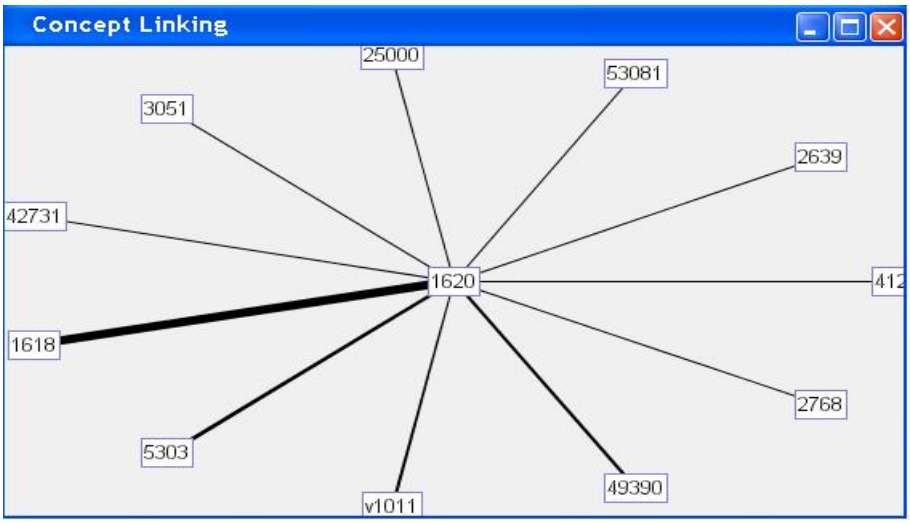
Table 3 shows the translations of these clusters. These code translations are provided at http://icd9cm.chrisendres.com/.

Table 3 Translation for the clusters

Cluster #	Description	Label
1	Unspecified Urinary tract infection, Pneumonia , Unspecified Anemia, Diabetes mellitus without mention of complication, Atrial fibrillation	Diabetes and Heart Problems
2	Diabetes mellitus without mention of complication, Coronary atherosclerosis, Unspecified Congestive heart failure, Old myocardial infarction	Diabetes and Heart Problems (CHF)
3	Unspecified Bronchus and lung, Pneumonia , Unspecified Congestive heart failure, Atrial fibrillation, Volume depletion	COPD and Heart problems
4	Tobacco use disorder, Upper lobe, bronchus or lung, Iatrogenic pneumothorax, Chronic airway obstruction, History of tobacco use	COPD and smoking
5	Depressive disorder, Pleura, Main bronchus, Esophageal reflux, Tobacco use disorder	Depression
6	Secondary malignant neoplasm of Bone, bone marrow, Brain and spinal cord , Hypopotassemia, Malignant neoplasm of Other parts of bronchus or lung,	Metastasizing Cancer
7	Tobacco use disorder, History of tobacco use, Secondary and unspecified malignant neoplasm of Intrathoracic lymph nodes, Malignant neoplasm of Lower lobe, bronchus or lung, Chronic bronchitis With (acute) exacerbation	COPD and cancer in the lymph nodes

We want to examine the relationship between lung cancer and other diseases. Hence, we use concept links of 1620; the links for 1622,1623,1624,1625,1628 and 1629 are similar.

Figure 5. Concept links for 1620, Malignant Neoplasm of Trachea



Note that most of the links are to code 1618 (shown with the widest line), Malignant neoplasm of Other specified sites of larynx. The other large links are to 5303 (Stricture and stenosis of esophagus), v1011 (Personal history of malignant neoplasm of Bronchus and lung), and 49390(Asthma).

Again, kernel density estimation was use to make a comparison of age, length of stay and cost by clusters. The average cost for cluster 6 is greater compared to other clusters. There is no big difference between clusters 1, 2, 3 and 7, which means that they have similar severity conditions. Cluster 5 has a slightly higher probability of a higher cost than cluster 4 (Figure 5).

Figure 5. Kernel Density Estimate for Total Charges by Clusters.

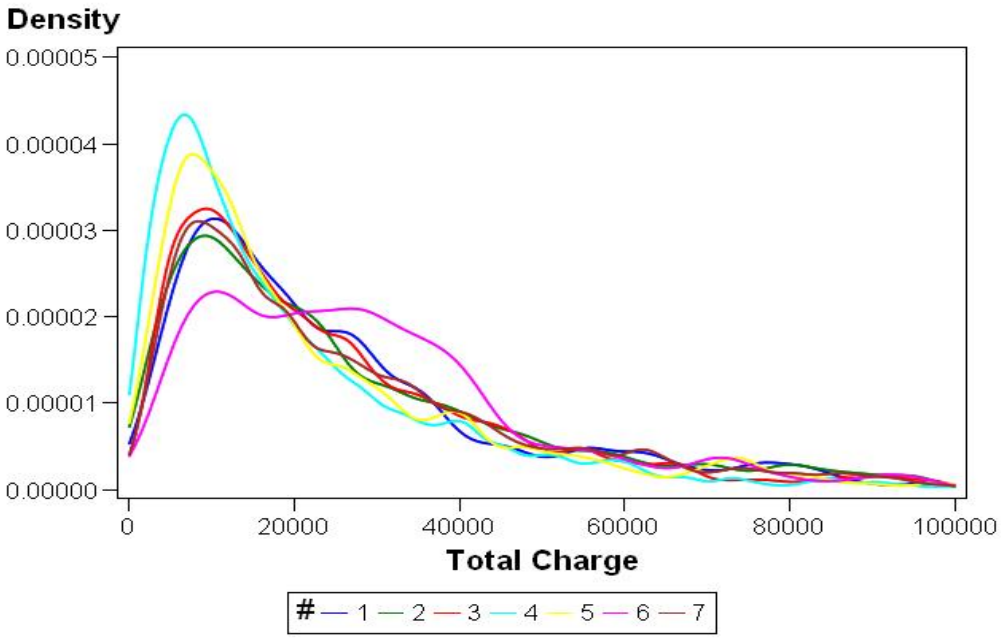
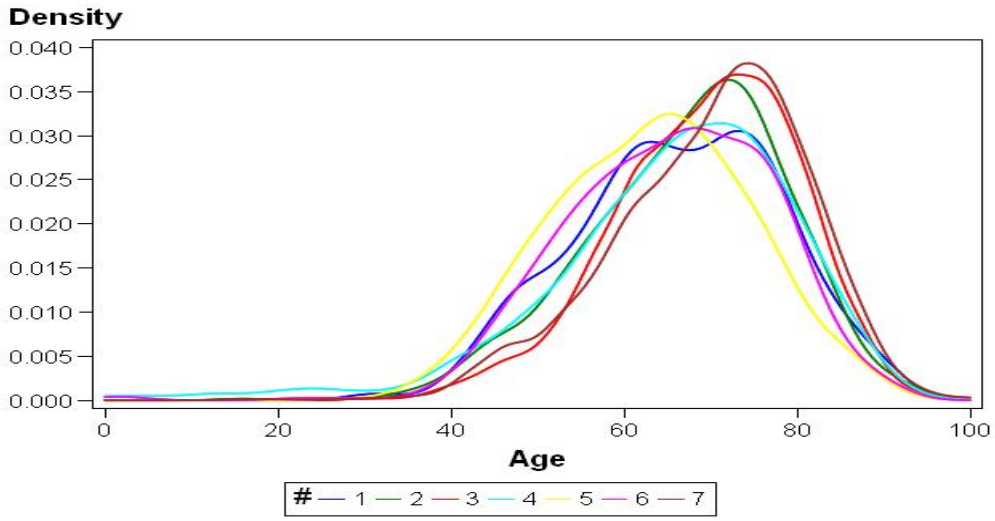
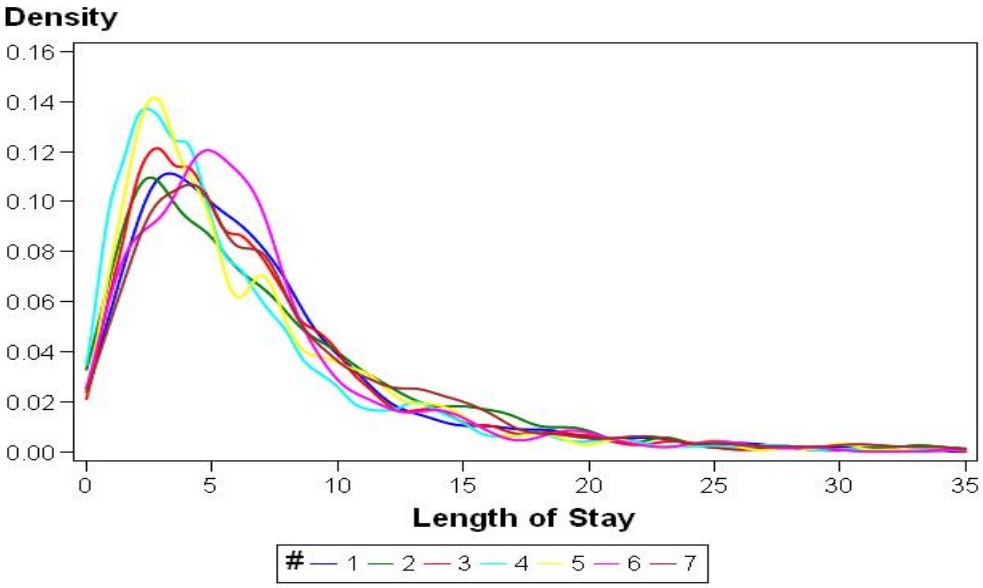


Figure 6. Kernel Density Estimate for Age by Clusters.



For the average age of each cluster, note that cluster 5 has the smallest average age, around 61, compared to other clusters. Clusters 1,4 and 6 have a similar average age of 70. Similarly, clusters 2,3 and 7 have an average age of 75.

Figure 7. Kernel Density Estimate for Length of Stay by Clusters.



Note that cluster 6 has a higher probability of a longer stay compared to the others. It would seem reasonable that patients at higher risk will stay longer and have higher cost.

CONCLUSION

Kernel Density Estimation was used to compare graphs that can be overlaid to give us more information. Here, we might conclude that older patients are more likely to have lung cancers that would lead to a higher probability of longer stay and higher costs for the treatment procedure. With text analysis on the diagnosis codes and KDE, it shows that malignant neoplasm of lobe, bronchus or lung is of higher risk and has a higher cost compared to other lung cancers.

REFERENCES:

[1] NIS; The NIS is part of the Healthcare Cost and Utilization Project (HCUP), sponsored by the Agency for Healthcare Research and Quality (AHRQ), formerly the Agency for Health Care Policy and Research. (http://www.ahrq.gov)

[2] Cerrito, Patricia, Data mining healthcare and clinical databases.

ACKNOWLEDGMENTS

Thanks to Dr. Patricia Cerrito, for aiding in the interpretation of the Text Miner results concerning ICD9 codes.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Guoxin Tang

Enterprise: University of Louisville

Address: Mathematics Department, University of Louisville

City, State, ZIP: Louisville, KY, 40292

Work Phone: (502) 435-3537

E-mail: guoxintang@hotmail.com