Paper D12-2009

Outcome Research for Diabetic Inpatients with SAS Enterprise Miner 5.2

Xiao Wang, Department of Mathematics,University of Louisville, KY

ABSTRACT

The main purpose of this paper is to evaluate and predict the diabetic inpatient outcomes in Medicare. In the study, we used data sets about inpatient claims, or beneficiary demography information for the year 2004, both of which come from the Chronic Condition Data Warehouse.

In this study, we used the Text Miner node to generate procedure and diagnosis clusters, preparing for kernel density estimation of the total charges, association analysis of the various procedures and prediction of the outcomes. We also used the link graphs and the rules table in the association analysis and different kinds of predictive models to analyze the outcomes. We utilized the CATX function to put all possible diagnosis or procedure codes into one text string .We also used the RXMATCH function, Random Sampling, SAS SQL and Base SAS.

Results show that many organ diseases and neurological disorders raise the costs of inpatient care. Although the expenditures on kidney disease are unexpectedly lower than those spent on diabetes itself, kidney disease has an important effect on the total charges, especially beyond 40,000 dollars. The procedures such as Hemodialysis and Angiocardiography are frequently used; most procedures related to cardiac diseases are utilized with other procedures. Another discovery is that only procedures and diagnoses are important in the prediction of mortality and total charges. The utilization day count is highly related to the total charges and conversely.

INTRODUCTION

In recent years, the diabetic inpatient care has consumed a high share of Medicare costs and there is growing pressure to utilize the scarce resources efficiently. Accordingly, patient outcomes have become an important focus of interests. The purpose of this paper is to provide useful information about patient outcomes to assist in improving healthcare management.

Diabetes is a chronic disease related to many organ dysfunctions as well as neurologic disorders. Statistics carried out by the American Diabetes Association show that heart disease strikes people with diabetes, twice as often as people without diabetes. It also points out that diabetes is the leading cause of new cases of blindness in people ages 20-74 and the cause of end-stage renal disease. The risk of a leg amputation is 15-40 times greater for a person with diabetes. Hence, it is essential to analyze organ diseases and nerve diseases in patients with diabetes. In diabetes,although the predominant treatment used is insulin, there are also many methodologies used to prevent and treat complications of diabetes(Davids.Bell, 2002). Hence, it is very necessary to analyze the importance of these procedures as well as the associations between them.

The Text Miner node in SAS Enterprise Miner can process volumes of textual data. After running this node, document clustering and concept linkage can be performed. Cluster analysis (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. In most cases, it is only a useful starting point for other purposes. Concept linkage connects related documents by identifying shared concepts between two unrelated data sets *(*Lavengood &Kiser, 2007). The purpose of the association node is to identify associations or relationships in the data that occur together or in a particular sequence. In this node analysis; two tools are often used: link graph and rules table. A link graph consists of various variables and the links represent the connections between the nodes. The bigger the node, the more important the variable; the heavier the line, the stronger the relationship. In a rules table, confidence, support and lift should be considered. Confidence is the proportion of times that the rule will contain the left side A and will also contain the right side B. Support is the proportion of times that the rule A and B occur together divided by the number of rules in the data set. A lift is the ratio of confidence divided by support. If a rule is of high confidence, support and lift, then it indicates a strong association.

METHOD

In this study, two data sets were used: one is Inpatient_base_claims, including 244,299 observations containing claim information for the year 2004; the other is Beneficiary _summar y _file covering 358,709 data containing beneficiary information for the year 2004. The variables to be used are as follows:

BENE_ID:	Encrypted 723 Beneficiary ID
BENE_SEX_IDENT_CD	Sex
BENE_RACE_CD	Beneficiary Race Code
BENE_AGE_AT_END_REF_YR	Age
BENE_ESRD_IND	End Stage Renal Disease indicator
ICD9_DGNS_CDn	Claim Diagnosis Code
CLM_TOT_CHRG_AMT	Claim Total Charge Amount
CLM_UTLZTN_DAY_CNT	Claim Utilization Day Count
NCH_DTNT_STATUS_IND_CD	NCH Patient Status Indicator Code
ICD9	AN abbreviation for the 9th edition of the International Classification of Diabetes and Related Health Problems

In the study, we first used Random Sampling in SAS Enterprise Guide to reduce the size of the data to 10,000; then, we joined sample data and the beneficiary data set by beneficiary ID. Next, we used kernel density estimation (KDE) to see how the total charges distributed among different races. The SAS code and KDE are shown below:

```
proc sort data=sasuser.ipclaimdemo
out=sasuser.sortipclaim;
by bene_race_cd;run;
proc kde  data=sasuser.sortipclaim;
univar clm_tot_chrg_amt/gridl=0 gridu=60000 method=snr
out=sasuser.kdesortipclaim; by bene_race_cd; run;
```

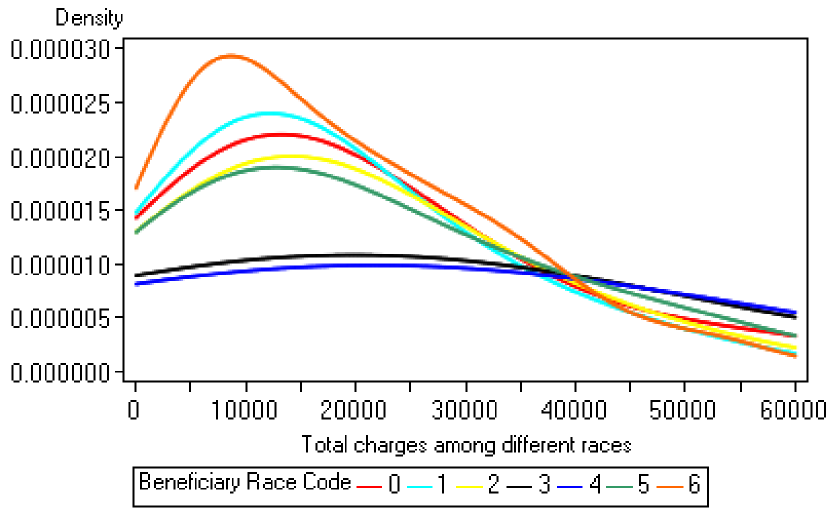


Figure1. KDE of total charges among different races

Figure 1 gives the estimates of the total charges. Before the value of 40,000 dollars occurs, the North American natives > the Whites>the Blacks>the Hispanics, and all of them is far greater than the Asians in terms of costs; while after that point, the Asians cost more than the other races.

Next, we want to see how organ dysfunctions and neurological disorders affect the total charges. We defined a string containing all possible diagnosis codes using the CATX statement, which concatenates character strings, removes leading and trailing blanks, and inserts separators. We used the following code:

```
data sasuser.ipclaim(keep=bene_id bene_sex_ident_cd  bene_race_cd
clm_tot_chrg_amt diagnoses );set sasuser.ipclaimdemo;
diagnoses=catx('',ICD9_DGNS_CD1,ICD9_DGNS_CD2,ICD9_DGNS_CD3,ICD9_DGNS_CD4,ICD9_DGNS_
CD5,ICD9_DGNS_CD6,ICD9_DGNS_CD7,ICD9_DGNS_CD8,ICD9_DGNS_CD9,ICD9_DGNS_CD10,ICD9_DGNS
_CD11,ICD9_DGNS_CD12,ICD9_DGNS_CD13,ICD9_DGNS_CD14,ICD9_DGNS_CD15,ICD9_DGNS_CD16);
run;
```

Now, we can input the new data into Enterprise Miner and use the Text Miner node, setting the default of number to Yes, Different parts of speech and Noun groups to No and the number of maximum clusters to 10. Then, we used Interactive-> Cluster documents to group the diagnoses. The results are displayed in Figure 2. In order to view how the clusters of diagnoses affect the total charges, we still used kernel density estimation. The SAS code is as follows:

```
libname emst "C:\Documents and Settings\Administrator\My
Documents\MySASFiles\9.1\EM_Projects\IPorganfailure\Workspaces\EMWS1";
data sasuser.ipclus(keep= _cluster_ _freq_ _rmsstd_ clus_desc);
set  emst.text_cluster;run;
data sasuser.iptchdem (keep= bene_sex_ident_cd bene_race_cd
clm_tot_chrg_amt diagnoses _cluster_);set emst.text_documents; run;
proc sort  data=sasuser.ipclus; by _cluster_;
proc sort  data=sasuser.iptchdem; by _cluster_;
data sasuser.ipkdetchdem;
merge  sasuser.ipclus sasuser.iptchdem; by _cluster_ ; run;
proc sort data=sasuser.ipkdetchdem out=sasuser.sortipkdetchdem;
by _cluster_  bene_sex_ident_cd ; run;
proc kde data=sasuser.sortipkdetchdem;
univar clm_tot_chrg_amt/  gridl=0 gridu=60000 method=snr out=sasuser.cluster; by
_cluster_  bene_sex_ident_cd ; run;
```

After running the SAS code, we get the KDE of total charges between the males and the females shown in Figure 3. The distributions of the costs for the male inpatients are different from the ones for the females. The first link graph in Figure 3 gives the relationships of the text cluster to male inpatient costs. From the shape of the graph, the clusters yield the relationships in terms of ordering; before the first cutpoint occurs at 19,200 dollars, cluster #5 is much greater than the other clusters; #1, #4 and #9 are almost the same and they are all greater than #2, 3. Clusters #6, 7, 8 are also almost the same, but they are all much smaller than the other clusters. Between the cutpoints 19,200 dollars and 33,000 dollars, the ordering is 1, 4, 5, 9>2, 3>6, 7, 8. After 33,000 dollars, there are no differences among all clusters. The graph for the female inpatients shows the relationships of the text clusters to the costs; in terms of ordering, we get 9>5>2>1,3>7.8>4,6 before the first cutpoint 10,650 occurs; between 10,650 and 16,800 dollars, 9>5>1>3>2>7,8>4,6. Between 16,800 and 19500 dollars, #1,3>2,5,7,8,9>4,6; and when the costs are above 34,650 dollars, cluster #6 is the greatest, and the clusters #1,2,3,4,5,7,8 are almost the same, but all of them are greater than cluster #9.

Clusters

# ▲	Descriptive Terms	Freq	Percentage	RMS Std.
1	27800, 29570, 25000, 3051, 2724	327	0.0327	0.0994078...
2	412, 41401, v4581, 41400, v4582	1414	0.1414	0.1202243...
3	4139, 42789, 2948, 41401, 2720	539	0.0539	0.1212629...
4	4019, 25000, 2449, 71590, 311	1949	0.1949	0.1284052...
5	25060, 36201, 25050, 3572, 2724	362	0.0362	0.0988086...
6	5849, 4280, 49121, 40391, 486	1828	0.1828	0.1266924...
7	4240, 25001, 4280, 4254, 42731	1276	0.1276	0.1260251...
8	3310, 5990, 2859, 2765, 486	1515	0.1515	0.1238352...
9	25000, 53081, 2724, 4019, 2720	790	0.079	0.1176632...

Figure 2. 9 Clusters of diagnoses

Cluster number	Diagnoses	Cluster label
1	Unspecified Obesity, Schizoaffective disorder, Diabetes mellitus without mention of complication, Tobacco use disorder, Other and unspecified hyperlipidemia	Diabetes
2	Old myocardial infarction, Of native coronary artery, Aortocoronary bypass status, Of unspecified type of vessel or native or graft, Percutaneous transluminal coronary angioplasty status	Heart disease
3	Other and unspecified angina pectoris, Other specified cardiac dysrhythmias, Other persistent mental disorders due to conditions classified elsewhere, Of native coronary artery, Pure hypercholesterolemia	Heart disease vascular disease
4	Unspecified Essential hypertension, Diabetes mellitus without mention of complication, Unspecified hypothyroidism, Osteoarthrosis which unspecified whether generalized or localized, Depressive disorder	vascular disease Diabetes

5	Diabetes with neurological manifestations, Background diabetic retinopathy, Diabetes with ophthalmic manifestations, Diabetes with ophthalmic manifestations, Other and unspecified hyperlipidemia	Ophthalmic disease Neurological disorder
6	Unspecified Acute renal failure, unspecified Congestive heart failure, Obstructive chronic bronchitis with exacerbation, Unspecified Hypertensive chronic kidney disease, Pneumonia	Heart disease Kidney disease
7	Mitral valve disorders, Diabetes mellitus without mention of complication, unspecified Congestive heart failure, Other primary cardiomyopathies, Atrial fibrillation,	Diabetes Heart disease
8	Alzheimer's disease, Urinary tract infection, unspecified Anemia, Volume depletion, Pneumonia	Others
9	Diabetes mellitus without mention of complication, Esophageal reflux, Other and unspecified hyperlipidemia, Unspecified Essential hypertension, Pure hypercholesterolemia	Diabetes vascular disease

Table 1. Translations for the 9 cluster

Male: **Female:**

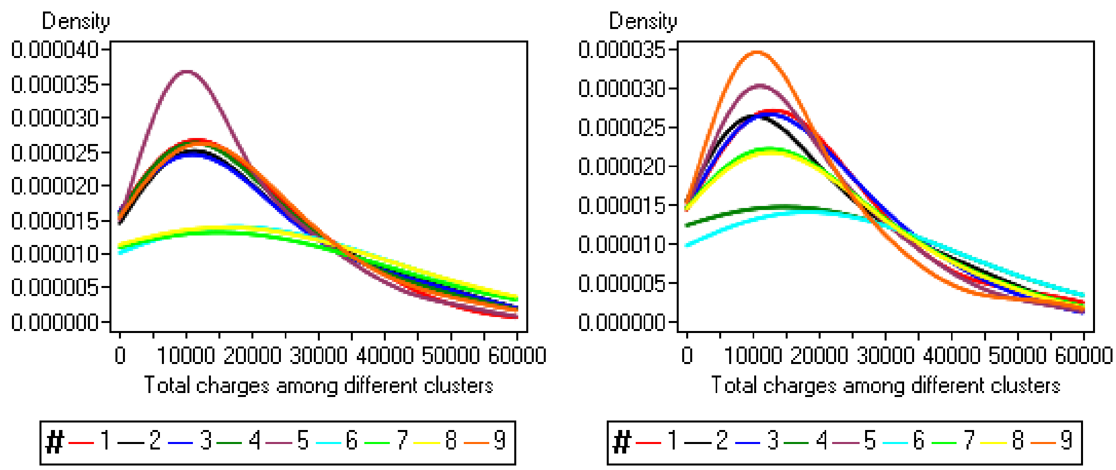
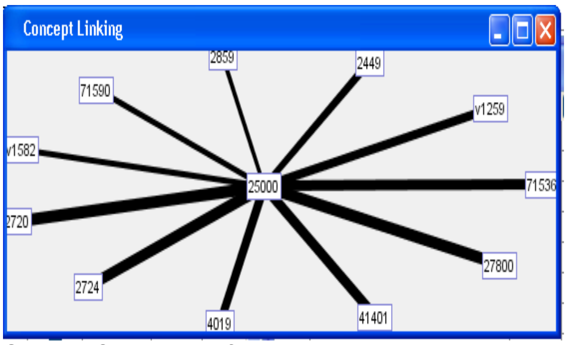
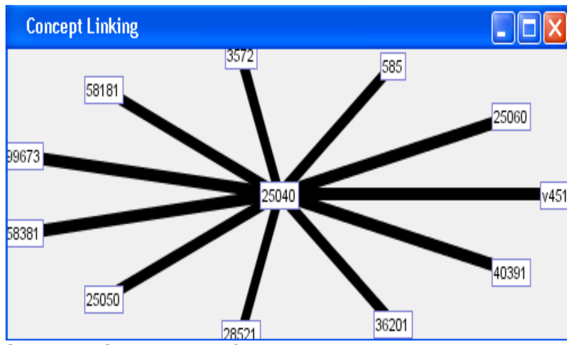


Figure 3. KDE of Total charges for diabetic inpatients by clusters

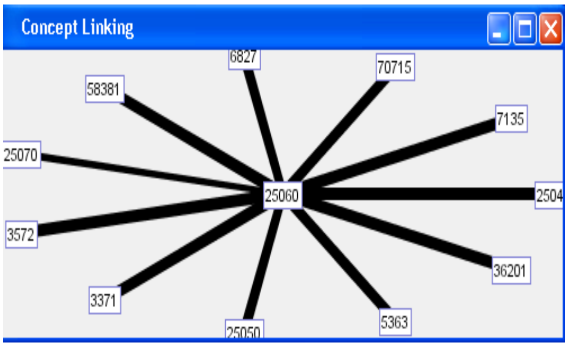
The cluster analysis just shows the costs by grouping the diagnoses; we need to see how organ diseases are related to diabetes. We used the concept link in Text Miner to show the relationships with the results displayed in Figure 4. The ICD9 codes that are not analyzed are translated in Table2.



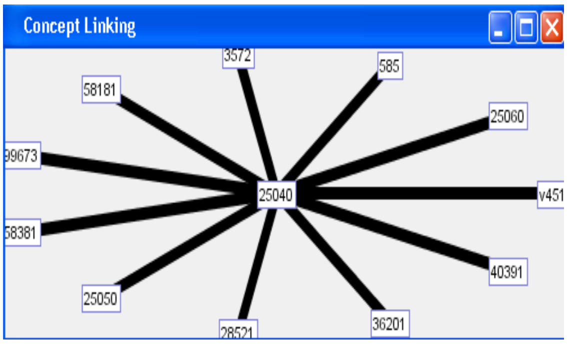
Output1. Concept link for 25000



Output 2. Concept link for 25040



Output3. Concept link for 25050



Output 4. Concept link to 25060

Figure 4. Linkages of organ diseases to diabetes

Output 1 shows the links to 25000 (diabetes mellitus without mention of complication). It indicates that the most prominent connections to diabetes are cardiovascular diseases such as 4019 (Unspecified Essential hypertension), 2720 (Pure hypercholesterolemia), 2724 (Other and unspecified hyperlipidemia). It also demonstrates that 41041(Coronary atherosclerosis of the native coronary artery, one kind of heart disease) has a strong connection to diabetes. Output 2 shows the links to 25040 (Diabetes with renal manifestations). It indicates that the larger links are to kidney diseases such as 58381(Nephritis and nephropathy), 40391(Unspecified Hypertensive chronic kidney disease) and 585 (Chronic kidney disease).The display in Output 3 shows that 36201 (Background diabetic retinopathy, one kind of eye disease) has the highest association with 25050 (Diabetes with ophthalmic manifestations).The other, larger links to 25050 are 25040, 25060, 3572(Polyneuropathy in diabetes) as well as 4039 (Unspecified Hypertensive chronic kidney disease).

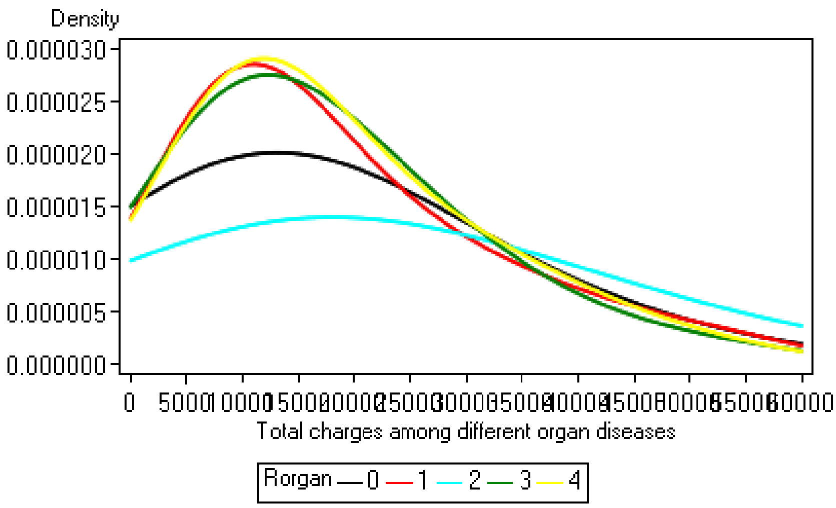
Output 4 shows that the most prominent connection to 25060 (Diabetes with neurological manifestations) is 3572. The other diseases closely related to 25060 are 3371(Peripheral autonomic neuropathy in disorders classified elsewhere), 36201(Background diabetic retinopathy), 25040 and 7135 (Arthropathy associated with neurological disorders).

2449	Unspecified hypothyroidism
27800	Unspecified Obesity
28521	Anemia in chronic kidney disease
2859	unspecified Anemia
36202	Proliferative diabetic retinopathy
5363	Gastroparesis
58181	Nephrotic syndrome in diseases classified elsewhere
6827	Other cellulitis and abscess :Foot(except toes)
70715	Ulcer of other part of foot
71536	Osteoarthrosis which not specified whether primary or secondary
71590	Osteoarthrosis which unspecified whether generalized or localized
99673	Due to renal dialysis device, implant, and graft
V1259	Other Diseases of circulatory system
V1582	History of tobacco use
V451	Renal dialysis status

Table 2. Translations for ICD9 diagnosis codes

All the results in Figure 4 indicate that diabetes is related to many diseases such as hypertension, blood diseases and so on, but we mainly focused on heart disease, kidney disease (including renal failure), ophthalmic disease and neurological disorders. We used the RXMATCH function to look for the initial code of ‘4280,’ which finds all inpatients with a diagnosis code related to heart disease and the same method for the other organ diseases. After that, we used IF, THEN statements to generate a new column. Then, we used the kernel density estimation to show the costs by organ diseases. The results are displayed in Figure 5.

```
data sasuser.iporgan(keep=CLM_ID CLM_TOT_CHRG_AMT CLM_DRG_CD
diagnoses Hea Kid Oculo Neu); set  sasuser.ipclaimdemo;
Hea=0;Kid=0;Oculo=0;Neu=0;
if(rxmatch('4280',diagnoses)>0)then Hea=1;…
data sasuser.organfailure;set sasuser.iporgan;if Hea=1 then Organ=1;if Kid=1 then
Organ=2;if Oculo=1 then Organ=3;if Neu=1 then Organ=4;run;
proc sort data=sasuser.rorgan out=sasuser.sortrorgan; by rorgan;run;
proc kde data=sasuser.sortrorgan;univar clm_tot_chrg_amt/gridl=0 gridu=60000
method=snr out=sasuser.kdesrorgan;by rorgan; run;
```



0: None of the organ diseases; 1:Heart diseases; 2:Kidney diseases;3: ophthalmic diseases, 4: Neurological disorders;

Figure 5. Total charges by different organ diseases

The graphs in Figure 5 indicate that before the costs reach the value of 9,900 dollars, the cost with heart disease, the cost with ophthalmic diseases and the cost with neurological disorders have almost the same probability, which is much higher than the cost without any of the organ diseases; and the probability for the cost with kidney disease is the smallest. However, after the cutpoint at 34,350 dollars, the density of the cost with kidney disease is higher than any other densities.Next, we need to study the procedures. We first used the following SAS code to rename the procedures and combine them into one column. We used this column for market basket analysis.

```
Data sasuser.ipproc1(keep=bene_id bene_sex _ident_cd prcdr);
set sasuser.ipclaimdemo;
prcdr=icd9_prcdr_cd1; where icd9_prcdr_cd1 ne' ';
data sasuser.ipproc2(keep=bene_id bene_sex_ident_cd
prcdr); set sasuser.ipclaimdemo;
prcdr=icd9_prcdr_cd2; where icd9_prcdr_cd2 ne' '; …
proc sql;
create table  sasuser.ipdemclm as
select * from sasuser.ipproc1  outer union corr
… select * from sasuser.ipproc6; run;
```

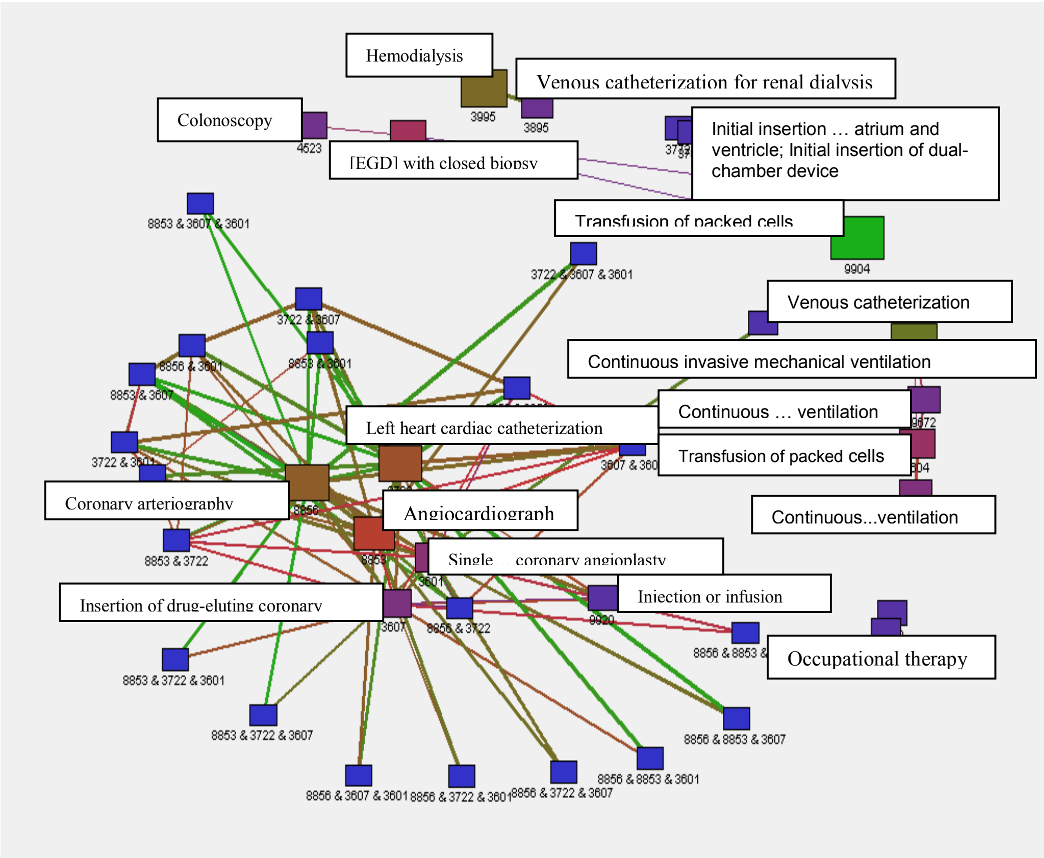


Figure 6. Link graph for ICD9

Code	Procedure
3601*	Single vessel percutaneous transluminal coronary angioplasty
3607*	Insertion of drug-eluting coronary artery stent(s)
3722*	Left heart cardiac catheterization
3772	Initial insertion of transvenous leads [electrodes] into atrium and ventricle
3783	Initial insertion of dual-chamber device
3893	Venous catheterization, not elsewhere classified
3895	Venous catheterization for renal dialysis
3995	Hemodialysis
4516	Esophagogastroduodenoscopy [EGD] with closed biopsy
4523	Colonoscopy
8853*	Angiocardiography of left heart structures
8856*	Coronary arteriography using two catheters
9339	Other physical therapy
9383	Occupational therapy
9604	Insertion of endotracheal tube
9671	Continuous invasive mechanical ventilation for less than 96 consecutive hours
9672	Continuous invasive mechanical ventilation for 96 consecutive hours or more
9904	Transfusion of packed cells
9920	Injection or infusion of platelet inhibitor

Table 3. Translations for important ICD 9 procedure code

Figure 6 shows all the major connections between different procedures. The procedures shown in table 1 are important since all of the rectangular boxes representing them are bigger than the others. Among the procedures, five of them are used for cardiac disease and one is related to hematic disease, which form 6 centers of the diagram; they were marked with an asterisk, '*' in table 3. The details about the 6 centers are discussed respectively in Figures 7 to 12. The output also indicates that Hemodialysis and Venous catheterization for renal dialysis are vital to the inpatients with diabetes, and there is a strong relationship between them. However, they have almost no connections to the other procedures.

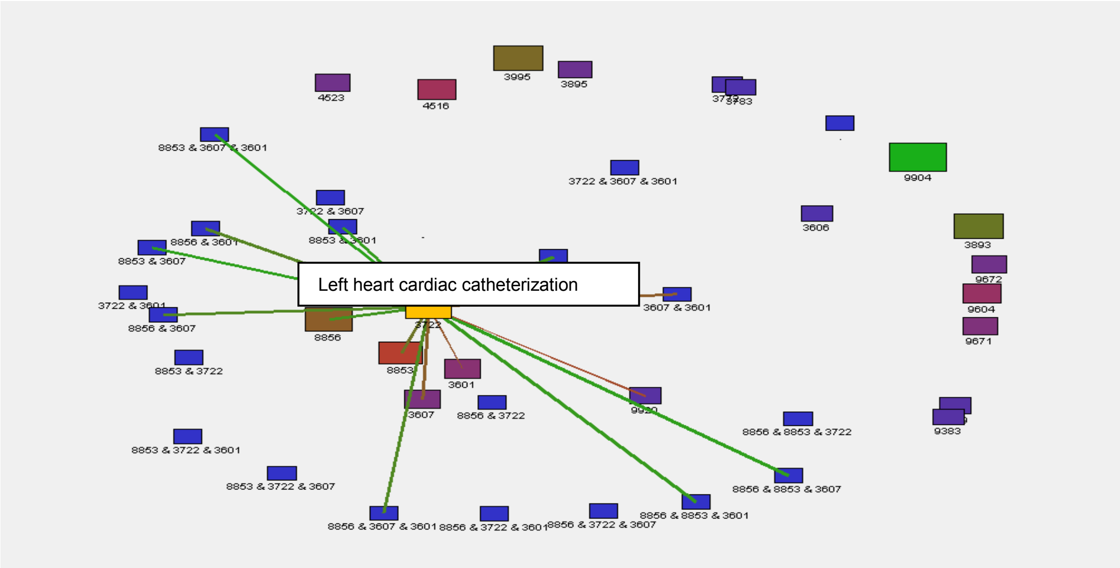


Figure 7. Link graph for Left heart cardiac catheterization

Figure 7 shows that left heart cardiac catheterization has strong relationships to the following procedures: Insertion of drug-eluting coronary artery stent(s), Angiocardiography of left heart structures, Coronary arteriography using two catheters; it is also strongly connected to the combinations with these procedures. It has weak relationships to Single vessel percutaneous transluminal coronary angioplasty and Injection or infusion of platelet inhibitor.

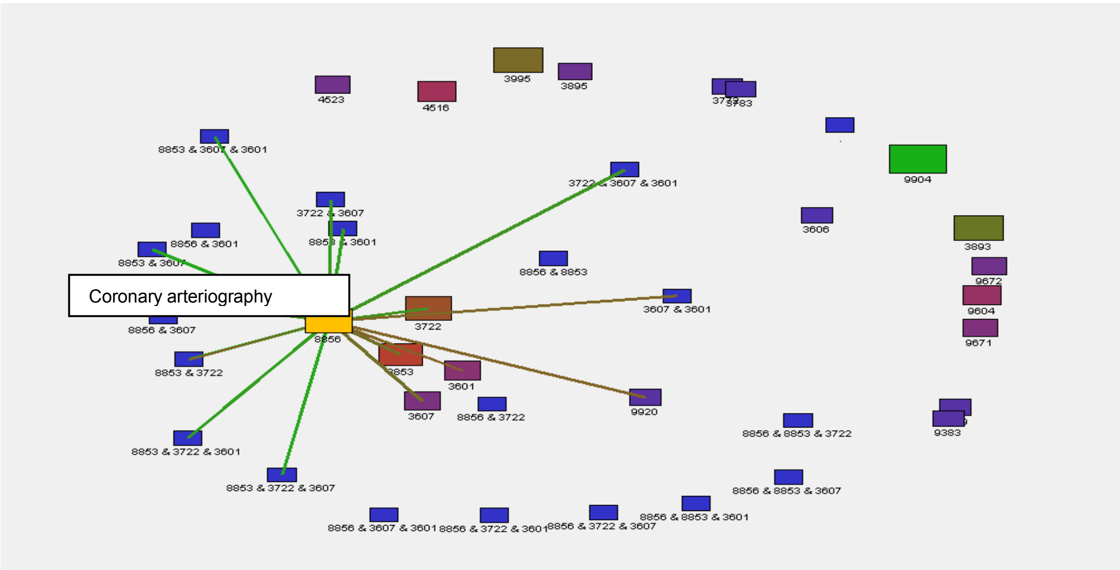


Figure 8. Link graph for Coronary arteriography using two catheters

Figure 8 demonstrates that Single vessel percutaneous transluminal coronary angioplasty, Insertion of drug-eluting coronary artery stent(s), Left heart cardiac catheterization, Angiocardiography of left heart structures as well as their combinations are strongly connected to Coronary arteriography using two catheters.

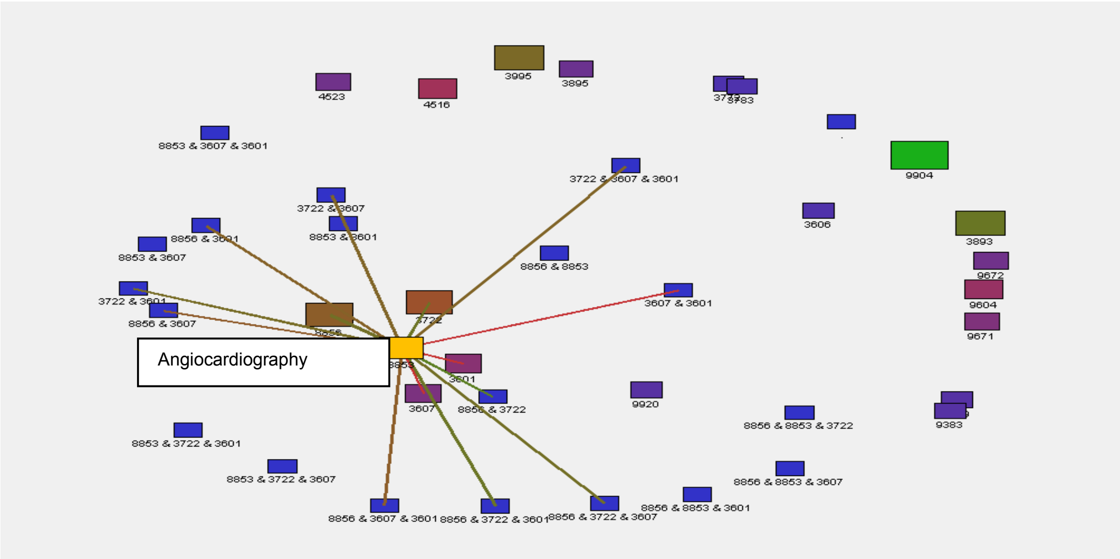


Figure 9. Link graph for Angiocardiography of left heart structures

Figure 9 shows that Angiocardiography has strong relationships with Left heart cardiac catheterization and Coronary arteriography and the combinations with them; it has weak connections to Single vessel percutaneous transluminal coronary angioplasty and Insertion of drug-eluting coronary artery stent(s) as well as their combinations.

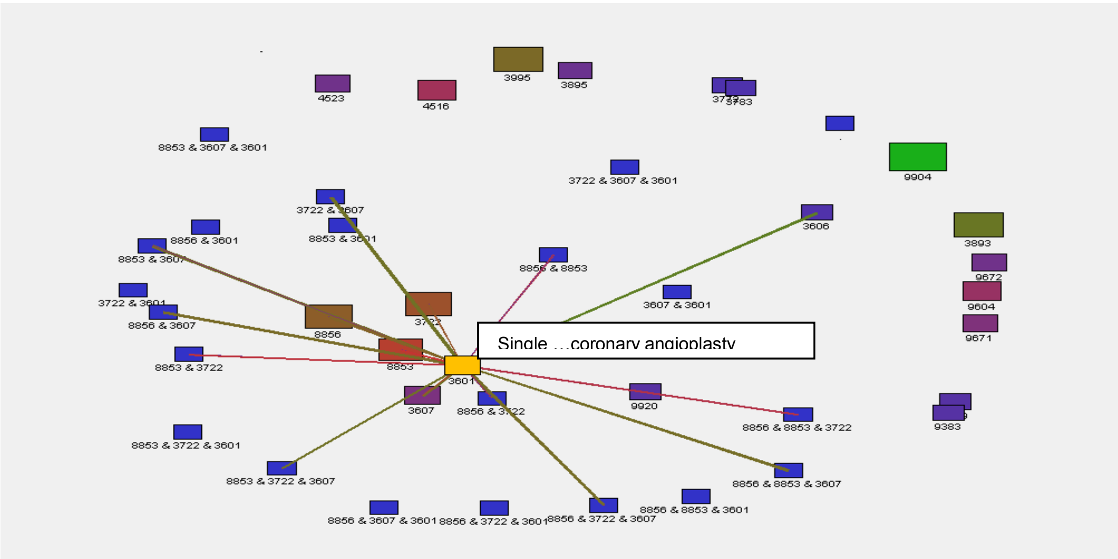


Figure 10. Link graph for Single vessel percutaneous transluminal coronary angioplasty

As figure 10 shows, Single vessel percutaneous transluminal coronary angioplasty is strongly related to the Insertion of a drug-eluting coronary artery stent(s) and its combinations as well as to the Insertion of a non-drug-eluting coronary artery stent(s). Its other connections to the other procedures are weak. In figure 11, Insertion of drug-eluting coronary artery stent(s) has a strong connection to Single vessel percutaneous transluminal coronary angioplasty, Left heart cardiac catheterization, and Coronary arteriography and their combinations; it has a weak relationship with Injection, infusion of platelet inhibitor and their combinations. In Figure 12, Injection or infusion of platelet inhibitor has a strong relationship to Coronary arteriography and weak relationships to the other procedures.

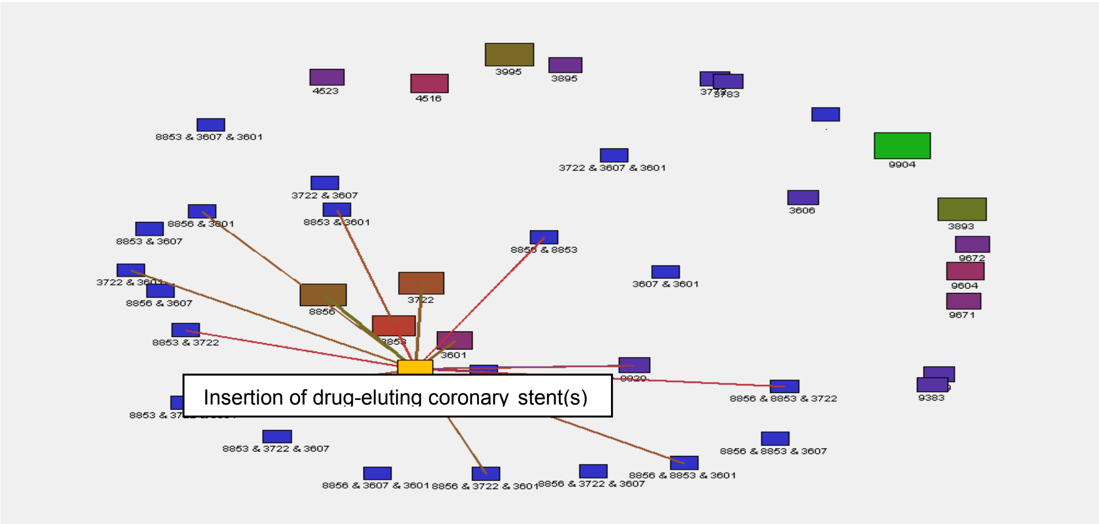


Figure 11. Insertion of drug-eluting coronary artery stent(s)

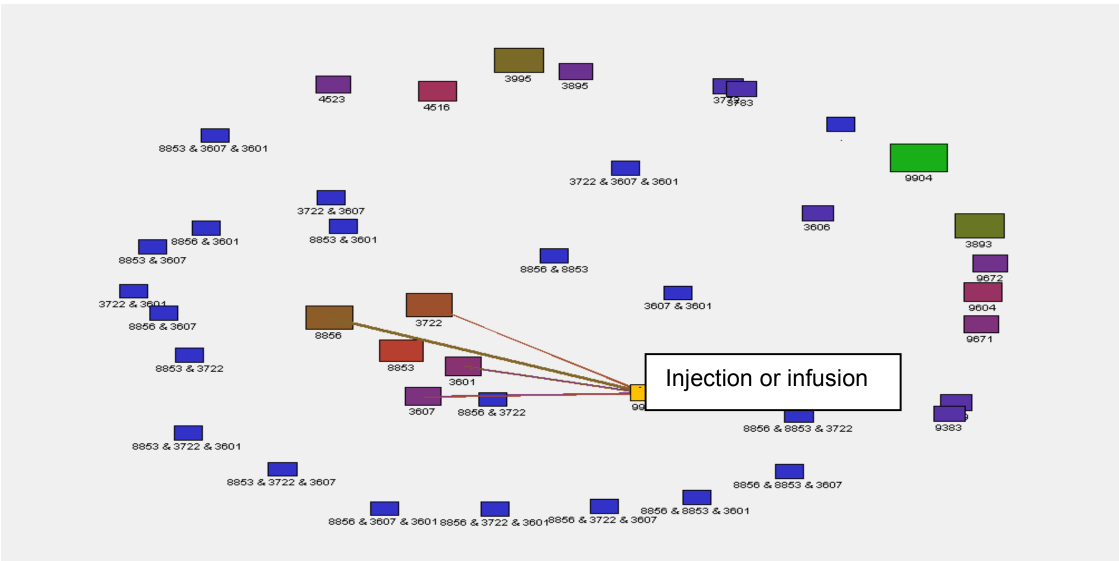


Figure 12. Link graph for injection or infusion of platelet inhibitor

The rules table is displayed in the descending order of confidence. In the table, many rules have high confidence and high lift, but they have low support. The Statistics line plot in Figure 13 also vividly shows that all rules have low support. A low support indicates that there is a small chance that both antecedent and consequence will be used at the same time. Hence, in the association node analysis, we mainly focused on the confidence and the lift. Table 5 just shows the important and meaningful rules since the rules with a low confidence or a low lift were discarded. We did not consider the combination cases since it is hard to decide which procedure in a combination is important. 100 per cent of confidence indicates that there is a high chance that the procedure, Initial insertion of transvenous leads [electrodes] into atrium and ventricle will be used given that the procedure, Initial insertion of dual-chamber device will be subsequently used. The lift value for this rule is 73.22, which indicates that the association between these two separate procedures is strong. For the same reason, the relationship between Occupational therapy and Other physical therapy is also strong. For the other rules in the table, their confidence values are higher, which indicates that it is very likely that subsequent procedures will be used if the antecedent procedure is used, since all the left confident values are high. However, the rules are not necessarily helpful because all their lift values are small.

Figure 13. Statistics line plot

Confidence (%)	Support (%)	Lift	Transaction Count	Rule
100	1.33	73.22	72	3772 ==> 3783
97.30	1.33	73.22	72	3783 ==> 3772
97.25	1.96	9.69	106	8853 & 3607 ==> 8856
97.09	1.85	9.67	100	8853 & 3722 & 3607 ==> 8856
96.43	1.50	9.60	81	8853 & 3607 & 3601 ==> 8856
95.24	1.48	10.20	80	8853 & 3607 & 3601 ==> 3722
94.72	6.96	9.43	377	8853 & 3722 ==> 8856
94.57	2.25	9.42	122	8853 & 3601 ==> 8856
94.50	1.90	10.12	103	8853 & 3607 ==> 3722
94.34	1.85	10.10	100	8856 & 8853 & 3607 ==> 3722
94.17	7.75	9.38	420	8853 ==> 8856
94.17	2.09	9.38	113	8853 & 3722 & 3601 ==> 8856
93.02	2.21	9.96	120	8853 & 3601 ==> 3722
92.86	2.40	9.25	130	3722 & 3607 ==> 8856
92.62	2.09	9.92	113	8856 & 8853 & 3601 ==> 3722
91.74	1.85	10.92	100	8853 & 3607 ==> 8856 & 3722
90.83	1.83	9.05	99	3722 & 3607 & 3601 ==> 8856
90.51	2.64	9.01	143	3722 & 3601 ==> 8856
89.92	8.40	8.96	455	3722 ==> 8856
89.76	6.96	9.61	377	8856 & 8853 ==> 3722

Table 4. Rules Table with Highest Confidence

Rules	Confidence (%)	Lift
Initial insertion … atrium and ventricle==> Initial insertion of dual-chamber device	100	73.22
Initial insertion of dual-chamber device==> Initial insertion … atrium and ventricle	97.28	73.22
Angiocardiography of left heart structures ==> Coronary arteriography	94.17	9.38
Left heart cardiac catheterization==> Coronary arteriography	89.92	8.96
Angiocardiography of left heart structures==> Left heart cardiac catheterization	89.24	9.56
Coronary arteriography==> Left heart cardiac catheterization	83.64	8.96
Hem dialysis== > Venous catheterization for renal dialysis	79.61	7.41
Insertion of drug-eluting coronary artery stent(s)==> Single… coronary angioplasty	78.68	18.37
Left heart cardiac catheterization==> Angiocardiography	78.66	9.56
Coronary arteriography==> Angiocardiography of left heart structures	77.21	9.38
Insertion of drug-eluting coronary artery stent(s)==> Coronary arteriography	77.16	7.08
Occupational therapy==> Other physical therapy	77.08	43.06
Other physical therapy==> Occupational therapy	76.29	43.06
Continuous invasive mechanical ventilation==> Colonoscopy	74.27	14.74
Injection or infusion==> Coronary arteriography	73.27	7.28
Single …coronary angioplasty==> Coronary arteriography	72.84	7.26
Insertion of drug-eluting coronary artery stent(s)==> Left heart cardiac catheterization	71.07	7.61
Single …coronary angioplasty==> Left heart cardiac catheterization	68.10	7.29
Single …coronary angioplasty==> Insertion of drug-eluting coronary artery stent(s)	66.81	18.37
Injection or infusion==> Left heart cardiac catheterization	64.36	6.89
Continuous invasive mechanical ventilation more than 96 hour==> Insertion of endotracheal tube	63.25	12.55
Continuous invasive mechanical ventilation more than 96 hour==>Continuous invasive mechanical ventilation	56.04	14.74
Single …coronary angioplasty==> Angiocardiography of left heart structures	55.60	6.75
Insertion of drug-eluting coronary artery stent(s)==> Angiocardiography	55.33	6.72
Insertion of endotracheal tube==> Continuous invasive mechanical ventilation more than 96 hour	38.46	12.55

Table 5. Confidence and lift for rules

Once we finished the analysis of diagnoses and procedures, we want to predict the outcomes. We used 0-1 indicators to define the new variable, mortality; we also defined the strings containing all possible diagnoses or procedures using the CATX function.

```
 data sasuser.ipmortal;  set sasuser.predictiveip;
if (NCH_PTNT_STATUS_IND_CD eq: 'B')then mortal=1; else mortal=0; run;
data sasuser.ipdiapro(keep=BENE_ID BENE_SEX_IDENT_CD BENE_AGE_AT_END_REF_YR
BENE_ESRD_IND DIAGNOSIS  PROCEDURE CLM_UTLZTN_DAY_CNT MORTAL CLM_TOT_CHRG_AMT);
set  sasuser.ipmortal;
diagnosis=catx('',ICD9_DGNS_CD1,ICD9_DGNS_CD2,ICD9_DGNS_CD3,ICD9_DGNS_CD4,ICD9_DGNS_
CD5,ICD9_DGNS_CD6,ICD9_DGNS_CD7,ICD9_DGNS_CD8,ICD9_DGNS_CD9,ICD9_DGNS_CD10,ICD9_DGNS
_CD11,ICD9_DGNS_CD12,ICD9_DGNS_CD13,ICD9_DGNS_CD14,ICD9_DGNS_CD15,ICD9_DGNS_CD16);
procedure=catx('',ICD9_PRCDR_CD1,ICD9_PRCDR_CD2,ICD9_PRCDR_CD3,ICD9_PRCDR_CD4,ICD9_P
RCDR_CD5,ICD9_PRCDR_CD6);run;
```

Next, we performed text mining again to cluster the diagnoses and procedures separately as shown in Tables 6 &7. Tables 6 & 7 list all diagnosis clusters and procedure clusters, and the important clusters are marked with an asterisk,'*', which will be discussed in the later model analysis. Most Diagnoses shown in Table 6 are related to heart disease, kidney disease, respiratory disease and vascular disease. Table 7 demonstrates that many diabetic inpatients need heart operations or eye operations.

#	Descriptive Terms	Freq
3	40391, 5990, 3572, 42731, 25040	39572
18	v4581, 42731, 41400, 41401, 4019	23258
2	2768, 2765, 5939, 5990, 2859	20514
1	25002, v1582, 25001, 2765, 4280	17981
7	5990, 2948, 78039, 496, 25000	17928
13	25000, 2724, 42789, 41401, 4019	17274
6	25000, 41400, 53081, 4019, 49121	16467
9	5849, 40391, 4280, 2765, 42731	11650
12	4019, 496, 2724, 25000, 53081	9704
8	6826, 49390, 78057, 4019, 27801	9373
15	5789, 2851, 2800, 56210, 25000	8204
10	496, 4240, 4280, 40391, 4254	8051
14	41400, 2724, 25000, 40391, 41401	7838
5	4019, 25000, 311, 2449, 53081	7826
16	25000, 4019, v1259, 41401, 2724	7209
19	29411, 5990, 2765, 29410, 3310	5844
4	53081, 73300, 4019, 2449, 71590	5807
17	v4365, v5789, 7812, 7993, 4019	5297
11	7318, 6827, 40391, 7854, 25060	4502

Table 6.Clusters of ICD 9 Diagnosis Code

#	Descriptive Terms	Freq
1	6495, 1642, 5317, 7956, 1364	108246
5	9604, 4513, 4311, 3893, 9672	19422
9	4542, 4516, 8152, 4443, 4525	19376
2	3601, 3723, 3606, 3722, 8856	17555
8	3324, 9339, 9383, 9671, 9604	16911
7	3772, 8872, 3895, 9904, 9671	16605
6	8411, 0309, 8848, 8154, 8622	15367
3	9394, 3812, 9921, 9390, 9929	14152
10	4525, 4573, 8753, 9907, 5459	9525
4	3942, 3943, 8609, 9749, 3995	7140

Table 7. Clusters of Procedures

Cluster #	ICD9 Diagnoses	Label
1*	Diabetes mellitus without mention of complication,History of tobacco use, Diabetes mellitus without mention of complication, Volume depletion , Unspecified Congestive heart failure	Heart disease
2*	Hypopotassemia, Volume depletion, Unspecified disorder of kidney and ureter,Urinary tract infection, Unspecified Anemia	Kidney disease Anemia
3*	Unspecified Hypertensive chronic kidney disease, Urinary tract infection, Polyneuropathy in diabetes, Atrial fibrillation, Diabetes with renal manifestations	Heart disease Kidney disease
4	Esophageal reflux, unspecified Osteoporosis, Unspecified Hypertensive chronic kidney disease, Unspecified hypothyroidism, Osteoarthrosis	Esophageal reflux Kidney disease Osteoarthrosis
5	Unspecified Hypertensive chronic kidney disease, Diabetes mellitus without mention of complication, Depressive disorder, Unspecified hypothyroidism,Esophageal reflux	Esophageal reflux Mental disorder
6*	Diabetes mellitus without mention of complication, Of unspecified type of vessel or native or graft , Esophageal reflux, Unspecified Essential hypertension, Obstructive chronic bronchitis with (acute) exacerbation	Hypertension
7	Urinary tract infection, Other persistent mental disorders, Other convulsions,Chronic airway obstruction, Diabetes mellitus without mention of complication	Mental disorder Respiratory disease
8	cellulitis and abscess(hand) , Asthma, Unspecified sleep apnea, Unspecified Hypertensive chronic kidney disease, Morbid obesity	Respiratory disease Kidney disease
9*	unspecified Acute renal failure, Unspecified Hypertensive chronic kidney disease,Congestive heart failure, Volume depletion, Atrial fibrillation	Kidney disease Heart disease
10	Chronic airway obstruction, Mitral valve disorders, Congestive heart failure,Unspecified Hypertensive chronic kidney disease, Other primary cardiomyopathies	Respiratory disease Heart disease
11	Other bone involvement in diseases, cellulitis and abscess(Foot), Unspecified Hypertensive chronic kidney disease, Gangrene, Diabetes with neurological manifestations	Diseases related to neurological manifestations
12	Unspecified Hypertensive chronic kidney disease, Chronic airway obstruction, Other and unspecified hyperlipidemia, Diabetes mellitus without mention of complication	Kidney disease Respiratory disease
13	Diabetes mellitus without mention of complication, hyperlipidemia, Other specified cardiac dysrhythmias, Of native coronary artery, Unspecified Hypertensive chronic kidney disease	Vascular disease
14	Of unspecified type of vessel, native or graft, hyperlipidemia, Diabetes mellitus without mention of complication, Unspecified Hypertensive chronic kidney disease, Of native coronary artery	vascular disease Kidney disease
15	Hemorrhage of gastrointestinal tract, Acute posthemorrhagic anemia, Secondary to blood loss, Diverticulosis of colon, Diabetes mellitus without mention of complication	Vascular disease
16*	Diabetes mellitus, Unspecified Essential hypertension, Other Diseases of circulatory system, Of native coronary artery, hyperlipidemia	Cardiovascular disease
17	Organ or tissue replaced by other means, Other specified rehabilitation procedure,Abnormality of gait, unspecified, Unspecified Hypertensive chronic kidney disease	Kidney disease
18	Aortocoronary bypass status, Atrial fibrillation, Of unspecified type of vessel or native or graft, Of native coronary artery, Unspecified Hypertensive chronic kidney disease	Heart disease Kidney disease
19	Dementia in conditions classified elsewhere with behavioral disturbance, Urinary tract infection, Volume depletion, Dementia in conditions classified elsewhere without behavioral disturbance, Alzheimer's disease	Mental disorders

Table 8. Translations for the 19 diagnosis clusters

Cluster #	ICD 9 Procedures	Label
1*	Insertion or replacement of non-inflatable penile prosthesis, Enucleation of eyeball with other synchronous implant, Bilateral inguinal hernia repair with graft or prosthesis,Open reduction of separated epiphysis, Discission of secondary membrane	Eye operation
2	Single vessel percutaneous transluminal coronary angioplasty, Combined right and left heart cardiac catheterization,Insertion of non-drug-eluting coronary artery stent(s),Left heart cardiac catheterization, Coronary arteriography using two catheters	Heart operation
3*	Vaccination against plague, Endarterectomy,Injection of antibiotic, Non-invasive mechanical ventilation,Injection or infusion of other therapeutic or prophylactic substance	Injection
4	Revision of arteriovenous shunt for renal dialysis, Removal of arteriovenous shunt for renal dialysis, Other incision of skin and subcutaneous tissue, Removal of other device from thorax, Hemodialysis	Renal Dialysis Hemodialysis
5*	Insertion of endotracheal tube, Other endoscopy of small intestine, Percutaneous [endoscopic] gastrostomy [PEG], Venous catheterization, Continuous invasive mechanical ventilation for 96 consecutive hours or more	Endotracheal tube catheterization
6	Amputation of toe, Other exploration and decompression of spinal canal, Arteriography of femoral and other lower extremity arteries, Total knee replacement, Excisional debridement of wound, infection, or burn	Lower extremity operation
7*	Initial insertion of transvenous leads [electrodes] into atrium and ventricle, Diagnostic ultrasound of heart, Venous catheterization for renal dialysis,Transfusion of packed cells, Continuous invasive mechanical ventilation for less than 96 consecutive hours	Heart operation
8*	Closed [endoscopic] biopsy of bronchus, Other physical therapy, Occupational therapy, Continuous invasive mechanical ventilation for less than 96 consecutive hours, Insertion of endotracheal tub	Some therapies
9	Endoscopic polypectomy of large intestine, Esophagogastroduodenoscopy [EGD] with closed biopsy, Partial hip replacement, Endoscopic control of gastric or duodenal bleeding, Closed [endoscopic] biopsy of large intestine	Intestine operation
10	Closed [endoscopic] biopsy of large intestine,Open and other right hemicolectomy,Intraoperative cholangiogram, Transfusion of other serum, Other lysis of peritoneal adhesions	Other operations

Table 9. Translations for the 10 procedure clusters

Once we defined the clusters, we used the following SAS code to generate two new variables: diacluster and procluster as well as a new data set.

```
data  sasuser.ipdiagnosis; set emws.text_documents;diacluster=_cluster_; run;
data  sasuser.ipprocedure; set emws.text2_documents;procluster=_cluster_; run;
proc sort data=sasuser.ipdiagnosis;   by BENE_ID;
proc sort  data=sasuser.ipprocedure; by BENE_ID; run;
data  sasuser.ippremortal  (drop= SVD_1-_SVD_100 PROB1-PROB19 _SVDLEN_ )
merge  sasuser.ipdiagnosis  sasuser.ipprocedure;
by  BENE_ID; run;
```

The following steps show the different kinds of models used to predict various targets.
First, we analyzed mortality and the diagram is displayed in Figure 14.

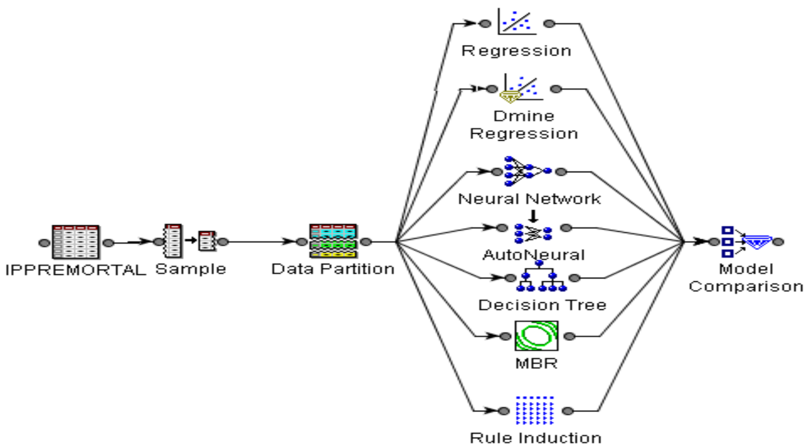


Figure 14. Predictive Models diagram

To predict mortality, we used the Regression model, the Dmine Regression model, the Neural Network model, the Auto Neural model, the Decision Tree model, the MBR model, the Rule Induction model and the Model Comparison model. Since mortality is a rare occurrence event, we changed Sample Method to Stratify, altered Stratify Criterion to Level based and changed Level Selection to Rarest level. Table 10 shows that the Model Comparison node identifies the decision tree as the optimal model. In this comparison, misclassification was used as the criterion to choose the optimal model. Although the misclassification rate of Rule Induction is a little higher than that of the Decision Tree model for the valid set; the other two misclassification rates of Rule are the lowest for the train set and validation set.

Selected Model	Model Node	Train: Akaike's Information Criterion.	Train: Average Squared Error.	Valid: Average Squared Error.	Test: Average Squared Error.	Train: Mis classification Rate.	Valid:Mis classification Rate.	Test:Mis classification Rate.
	AutoNeural	12427.2102	0.2629	0.2629	0.2629	0.5	0.5	0.5
	DmineReg	NaN	0.17601	0.1778	0.1759	0.2634	0.2623	0.2646
	MBR	-7091.3298	0.2174	0.2446	0.2467	0.3614	0.4304	0.4327
	Neural	9159.4143	0.1760	0.1823	0.1795	0.2668	0.2748	0.2726
	Reg	9083.0360	0.1771	0.1788	0.1764	0.2664	0.2654	0.2629
Y	Rule	NaN	NaN	NaN	NaN	0.2528	0.2640	0.2596
	Tree	NaN	0.1809	0.1856	0.1831	0.2577	0.2601	0.2637

Table 10. Fit statistics of the comparison model targeting at mortality

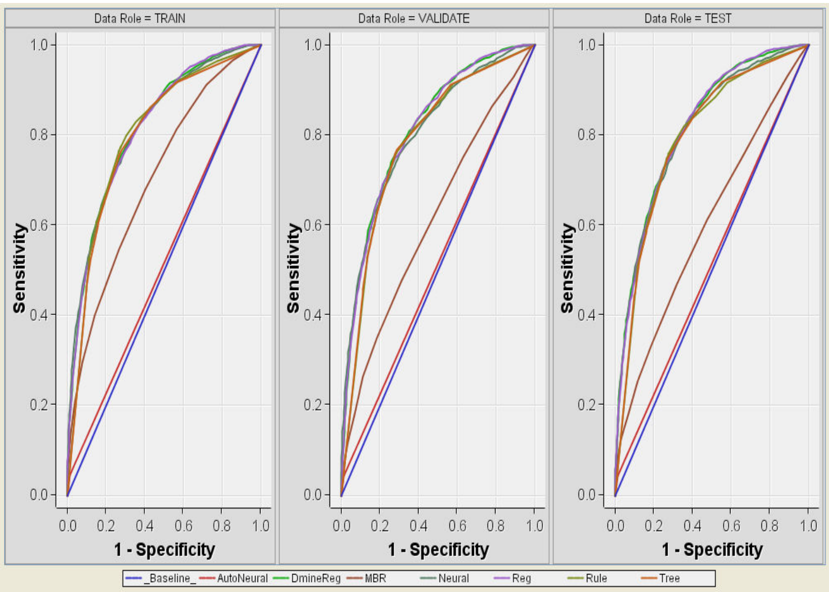


Figure 15.Roc Chart for Mortality prediction

The roc maps in Figure 15 show that for all the three sets: Train, Validate and Test, there are no big differences in accuracy among the various models.The lift for a given decile is the ratio of the target density for the decile to the target density over all the test data. Random chance is represented by the lift value1.0 and the lift value higher than 1.0 indicates a higher level of prediction. Therefore, according to the lift curves, we can find the patients at highest risk of dying. Figure 16 demonstrates that except for the MBR node, there are no differences among the other nodes in terms of the prediction of mortality. In the train set, validate set and test set, 40 per cent of beneficiary records have a higher level of prediction than just chance.

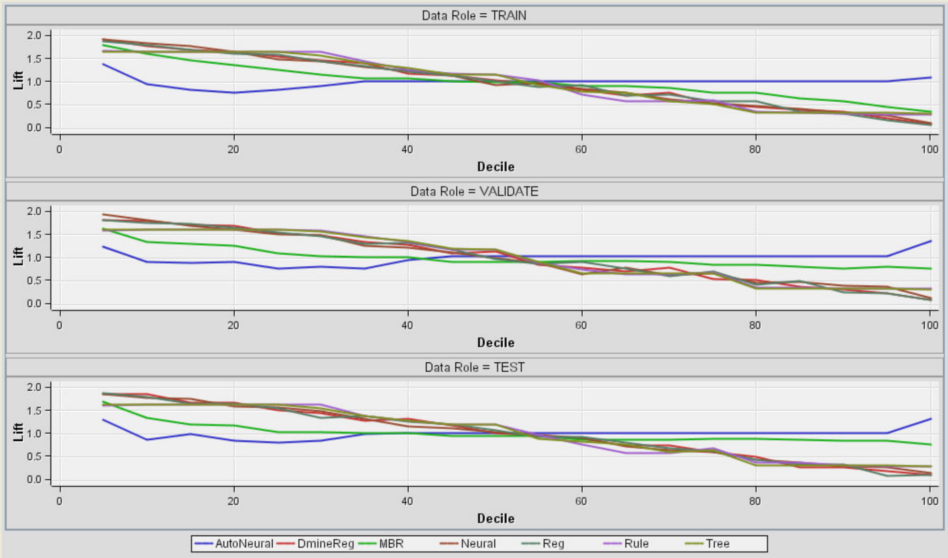


Figure 16.Lift curve for predictive model for mortality

After comparison, the Rule Induction model is still the best one. We analyzed the Tree Decision model instead, since its misclassification rates are only a little higher than those of the Rule and we can define a more meaningful model. The results of the Tree are displayed in Tables 11 and 12, and Figure 17.

Target	Statistics label	Train	Validation	Test
Motal	Misclassification Rate	0.25769	0.2600	0.2637
Motal	Average Squared Error	0.1809	0.1856	0.1831

Table 11.Fit statistics of Tree targeting at mortality

NAME	IMPORTANCE	VIMPORTANCE	RATIO
PROCLUSTER	1.0000	1.0000	1.0000
DIACLUSTER	0.7254	0.6807	0.9384
Age	0.3845	0.3993	1.0386
Utilization Day Count	0.3364	0.3464	1.0297
Total Charge	0.2466	0.2528	1.0252

Table 12.Variable Importance in Tree targeting at mortality

Table 11 shows that both the misclassification rate and average squared error are low; hence, this model is relatively good. Results in Table 12 demonstrate that the order of importance in the levels is Procedures > diagnoses > age > utilization day count > total charges. Next, the tree diagram will display how the input variables affect mortality.

In the tree diagram shown in figure 17, the first segment is divided on the procedure cluster, indicating that procedures are essential to the mortality; the next split is based upon the diagnosis cluster. The following split criteria vary from the left side to the right side. Age has no relation to mortality related to the procedure cluster #5 (endotracheal tube and catheterization) and #7 (Some heart operations); on the left side, age is an important variable. Before the age of 82.5, both total charges and utilization day count should be considered, while after that, only utilization day count should be focused on. Next, we will examine the total charges. Since the total charges form an interval variable, we did not use the Rule Induction model.

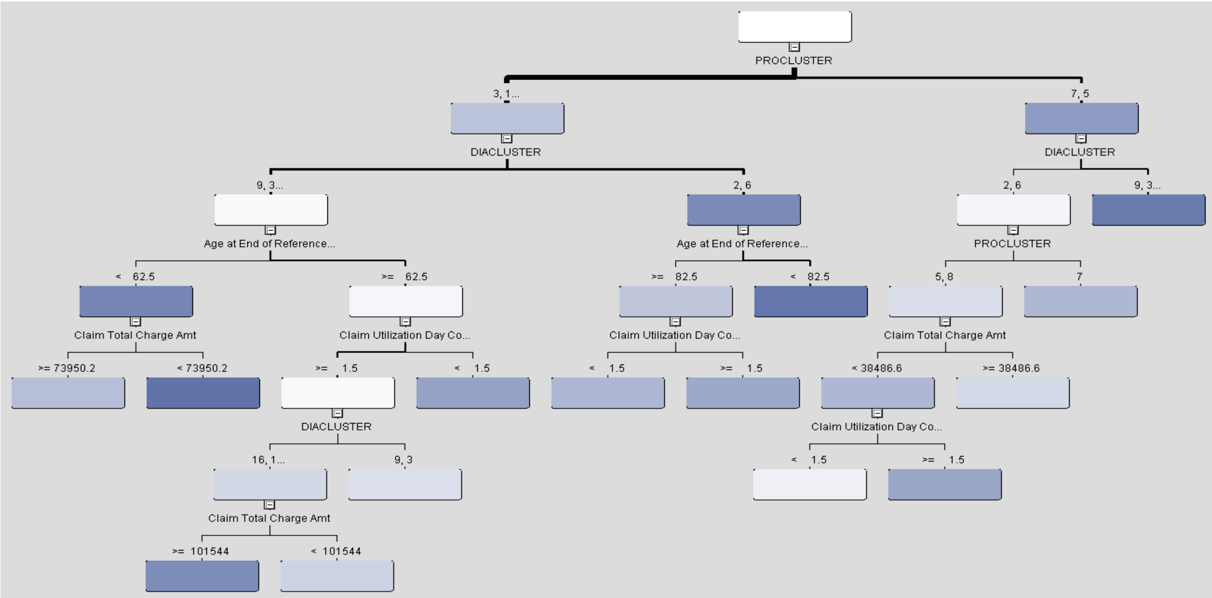
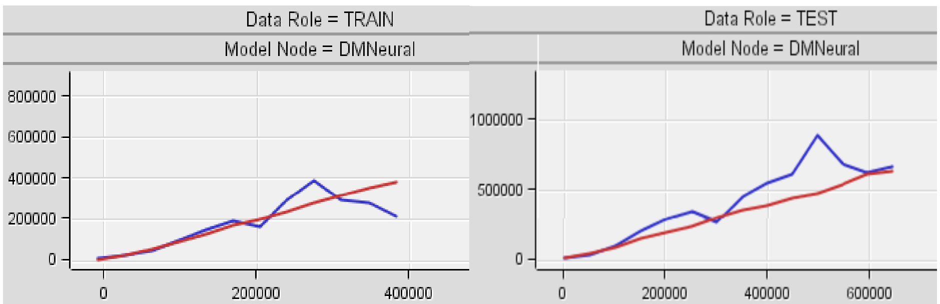


Figure 17. Tree diagram aiming at mortality

Selected Model	MODEL	Train: Akaike's InformationCriterion.	Train:Average SquaredError.	Valid: Average Squared Error.	Test: Average Squared Error.
	Tree3	NaN	7.61E+08	1.28E+09	1.81E+09
	Reg2	200721.4565	8.27E+08	1.21E+09	1.77E+09
	Neural2	207283.3272	1.60E+09	2.02E+09	3.00E+09
	MBR2	201426.456	8.95E+08	1.61E+11	1.62E+11
	DMNeural	201966.7255	9.42E+08	1.21E+09	1.82E+09
Y	DmineReg2	NaN	7.87E+08	1.23E+09	1.67E+09
	AutoNeural2	208794.5095	1.87E+09	2.31E+09	3.52E+09

Table 13. Fit Statistics of comparison targeting at total charge

The Comparison node automatically selects the Dmine Regression as the optimal model since its average squared errors are relatively small for the train set, validate set and test set.



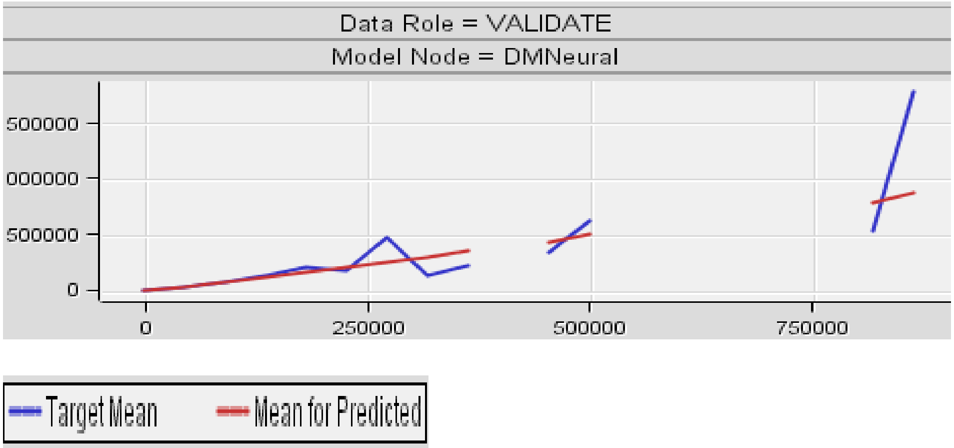


Figure18. Details of predicted means for DMine Regression Model

Figure 18 shows that the DMine Regression Model gives a relatively precise prediction of the means for both the train set and test set, but it provides a poor prediction for the validate set. After comparison, the DMine Regression is still the best model and we examined the details.

Effect	DF	R-square	F value	p –value	sum of square	Error Mean square
Var: UTLZTN_DAY_CNT	1	0.4857	9228.2991	<.0001	8.83E12	56995870
Class: PROCLUSTER	9	0.0606	144.9769	<.0001	1.10E12	844932727
Class: DIACLUSTE	18	0.0057	6.9099	<.0001	103957349835	835823770

Table14. Effects Chosen for Target: Total Charges

According to the r-square in table 14, the utilization day count can account for almost 50 per cent of the variability of the total charges while the procedures explain only 6 per cent and the diagnoses have nothing to do with the variability of the charges.

Effect	DF	R-Square	Sum of squares
Model	28	0.552098	1.0037863E13

Table 15.The Final ANOVA Table for Target: Total Charges

Table15 demonstrates that the input variables, utilization day count, procedures, and age can account for 55 percent of the variability of the total charges; hence, the predictive model is good.

Finally, we studied the utilization day count. Since it is also an interval variable, we utilized the same models as we did for the total charges. After comparison, the Decision Tree model is found to be the optimal model shown in table 16 and we will discuss it in detail.

Selected Model	Model	Train: Average Squared Error	Valid: Average Squared Error	Test: Average Squared Error
	AutoNeural5	[illegible].5422	49.4181	55.1114
	DMNeural4	29.3692	48.6671	48.7307
	DmineReg5	25.4773	34.2649	40.1687
	MBR5	26.3273	33.3339	32.8590
	Neural5	31.0633	34.1879	37.7437
	Reg5	25.7261	31.9023	33.7555
Y	Tree4	24.2618	30.0026	28.9996

Table 16. Fit statistics of comparison node targeting at utilization day count

Obs	NAME	IMPORTANCE	VIMPORTANCE	RATIO
1	Claim Total Charge Amt	1.00000	1.00000	1.00000
2	PROCLUSTER	0.25459	0.24900	0.978043
3	DIACLUSTER	0.22930	0.20885	0.91085

Table 17. Variable Importance of Decision Tree

Table 17 indicates the importance level of the input variables among all the variables; the total charges are decisive to the utilization day count, and the other two prominent variables are the procedures and the diagnoses.

CONCLUSION

After the study, we concluded that many organ diseases and neurological disorders indeed have important effects on the costs of inpatients with diabetes. Heart diseases, eye diseases and neurologic problems raise the inpatients costs. Although the expenditures on kidney diseases is unexpectedly lower than the ones on diabetes itself, after the total charges reach 34,350dollars, kidney disease begins to increase the inpatient costs. Hence, to reduce the costs, all inpatients with diabetes should pay more attention to their kidney disease, and to use prevention to avoid kidney disease. We also discovered that there are several procedures such as Hemodialysis and Angiocardiography that are prominent for diabetic inpatients. Among the various procedures, the ones utilized for cardiac disease treatments are related to many different procedures. Association analysis also shows that Hemodialysis is strongly related to Venous catheterization for renal dialysis. Another discovery is that neither age nor the end stage renal disease is the key factor to mortality, which is contrary to widely held belief. We also discovered that both the procedures and the diagnoses are important in the prediction of mortality and the total charges. The utilization day count plays a vital role in predicting the total charges, and the latter is also prominent to the utilization day count. However, there is still much left to do in the future. For instance, we will need to examine exactly which procedures and which diagnoses are vital to the prediction. We also need to study the other factors that affect the total charges.

REFERENCES

American Diabetes Association. *Diabetes and Cardiovascular (Heart) Disease* http://wwww.diabetes.org/diabetes-statistics/heart-disease.jsp

Anonymous, *Cluster Analysis*, http://www.statsoft.com/textbook/stcluan.html

Davids. H. Bell, Chronic Complications of Diabetes *Southern Medical Journal* ,2002. http://www.medscape.com/viewarticle/426919

Cerrito, P. (2008) *Data Mining Healthcare and Clinical Database*

Cerrito, P. (2006) *Introduction to Data Ming Using SAS Enterprise Miner*, SAS Press.

Kathryn A. Lavengood and Pam Kiser. Information Professionals in the Text Mine. *ONLINE, Vol.31, No. 3, May/June 2007.* http://www.infotoday.com/online/may07/Lavengood_Kiser.shtml

Matignon, R (2007) *Data Mining Using SAS Enterprise Miner*, John Wiley & Sons, Inc.

ACKNOWLEDGMENTS

We appreciate the use of the CCW Data Warehouse from the Centers for Disease Control. Thanks to my advisor, Dr. Patricia Cerrito, for aiding in SAS code and interpretation of some results in this paper.

CONTACT INFORMATION

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

Name: Xiao Wang
Enterprise: University of Louisville
Address: 2301 S 3Rd Street
City, State ZIP: Louisville, KY, 40292
E-mail: x0wang16@louisville.edu.