Paper 244-2008

# Text and Data Mining to Investigate Expenditures on Prescribed Medicines

Guoxin Tang, University of Louisville, Louisville, KY


## ABSTRACT

This Prescribed Medicines File from MEPS (Medical Expenditure Panel Survey) provides detailed information on household-reported prescribed medicines for a nationally representative sample of the civilian, non-institutionalized population of the United States and can be used to make estimates of prescribed medicine utilization and expenditures. For this study, we focused on the total cost, self-pay, Medicare cost, Medicaid cost and private insurance. It is the purpose of this study to determine the relationship between medication and patient condition, and to examine cost and patient condition in the analysis of these data. We use Text Miner to examine combinations of medications used in relationship to patient therapeutics. In addition, we will use time series methods to investigate costs in relationship to patient therapeutics.

The way to approach this purpose is to define the patient condition codes as a text string with the Text miner node. Then the terms window gives us the frequency and number of records for each condition. By clustering the text strings, the results show us the weights of the patient self-pay, Medicare cost, Medicaid cost and private insurance in the total cost. In addition, we used some time series features in Enterprise Guide to create time series data. Here, we accumulated using the average. There are a total of 317,076 observations at the beginning. After accumulating times by month, the number of records decreased to 178. Then we used the explore node in SAS Enterprise Miner to explain the weights of the other four parts in the total cost.


## INTRODUCTION

The MEPS Household Component uses questionnaires to individual household members and their medical providers to collect nationally representative data on demographics, health conditions, health status, use of medical care services, charges and payments, access to care, satisfaction with care, health insurance coverage, income, and employment.

This public use file provides detailed information on household-reported prescribed medicines for a sample of the civilian, non-institutionalized population of the United States and can be used to estimate costs and use of medical services. Each record represents prescribed medicine that was purchased during a calendar year by an individual, and by household. The prescribed medicine was reported as being purchased or otherwise obtained by the household and includes the following: an identifier for each unique prescribed medicine, detailed characteristics associated with the event (e.g., national drug code (NDC), medicine name, etc.), patient conditions, if any, associated with the medicine, the date on which the person first used the medicine, costs and sources of payments, types of pharmacies that filled the household's prescriptions, and a weighting factor.

## METHODS

In this paper, there were nine years of data, 1996 to 2004, which included 317,076 observations. Here we focused on the variables RXNAME, PREGCAT, TC1, TC1S1, TC1S1_1, TC1S1_2, TC1S2, TC1S2_1, TC2, TC2S1, TC2S1_1, TC2S1_2, TC2S2, TC2S2, RXSF04X, RXMR04X, RXMD04X, RXPV04X and RXXP04X where

| | |
|---|---|
| RXName | Medication Name |
| PREGCAT | Multum Pregnancy Category. |
| TC1 | Multum therapeutic class #1. |
| TC1S1 | Multum therapeutic sub-class#1 for TC1. |
| TC1S1_1 | Multum therapeut sub-sub-class for TC1S1. |
| TC1S1_2 | Multum therapeut sub-sub-class for TC1S1. |
| TC1S2 | Multum therapeutic sub-class#2 for TC1. |
| TC1S2_1 | Multum therapeut sub-sub-class for TC1S2. |
| TC2 | Multum therapeutic class #2. |
| TC2S1 | Multum therapeutic sub-class#1 for TC2. |
| TC2S1_1 | Multum therapeut sub-sub-class for TC2S1. |
| TC2S1_2 | Multum therapeut sub-sub-class for TC2S1. |
| TC2S2 | Multum therapeutic sub-class#2 for TC2. |
| RXSF04X | Amount paid, self or family . |
| RXMR04X | Amount paid, Medicare. |
| RXMD04X | Amount paid, Medicaid. |
| RXPV04X | Amount paid, private insurance. |

| | |
|---|---|
| RXXXP04X | Sum of payment . |
| PREGCAT | Pregnancy category variable - identifies the FDA pregnancy category to which a particular drug has been assigned. Multum Pregnancy Category. |
| TCn | Therapeutic classification variable - assigns a drug to one or more therapeutic/chemical categories; can have up to three categories per drug. |
| TCn_TCnSn | Therapeutic sub-classification variable - assigns one or more sub-categories to a more general therapeutic class category given to a drug. |
| TCnSn_n | Therapeutic sub sub-classification variable - assigns one or more sub subcategories to a more general therapeutic class category and sub-category given to a drug. |

First, we used SAS Enterprise Guide to preprocess the data set. Since we want to examine combinations of medications used in relationship to patient therapeutics, the way to work with all the variables is to bring all of them into one column as a string of codes, using the CATX function, which concatenates character strings, removes leading and trailing blanks, and inserts separators. The expression below shows an example of using this function to create a text string for the procedure codes.

```
CATX(‘ ’,medications.PREGCAT ,medications.TC1 ,medications.TC1S1,
medications.TC1S1_1 ,medications.TC1S1_2,
medications.TC1S2,medications.TC1S2_1 ,medications.TC2 ,
medications.TC2S1 ,medications.TC2S1_1 ,medications.TC2S1_2,
medications.TC2S2 )
```

This expression makes a string of codes from Column PREGCAT to TC2S2 with a space between them. We also make a string of codes with RXNAME and Condition, called Medi_Cond. Then we filtered the data set such that there were no missing values in the column, Medi_Cond. As a result, the number of observations reduced to 202,516. Then the filtered data were analyzed by using the Text Miner node in Enterprise Miner.

Time series models were also used on the Prescribed Medicines File data to analyze the total cost, self-pay, Medicare cost, Medicaid cost and private insurance. We used some time series features in Enterprise Guide to create time series data. Here, we accumulated using the average. There are a total of 317,076 observations at the beginning. After accumulating, the number of records decreased to 178.

## RESULTS

We use the dataset, including the concatenated columns defined using the CATX function, in SAS Enterprise Guide. Then we try to find the relationship between Medication and Conditions. Figure 1 shows the medications according to Condition=A 115 119 -1 -1 -1 -1 -1 -1 -1 for one patient. Here, Condition=A 115 119 -1 -1 -1 -1 -1 -1 -1 is a text string for many patients who had the same situation. This condition meant that controlled studies showed no risk for Pregnancy; Nutritional products were used for therapeutic class #1, Vitamins were used for sub-class#1 and the therapeutic classes for TC1S1_1, TC1S2, TC2, TC2S1, TC2S2, and TC3 were inapplicable. We know that the medications focused on PYRIDIXINE and VITAMIN B6. Therefore, we have the combination of medications for this Condition.

**Figure 1. The medications according to Condition=A 115 119 -1 -1 -1 -1 -1 -1 -1**

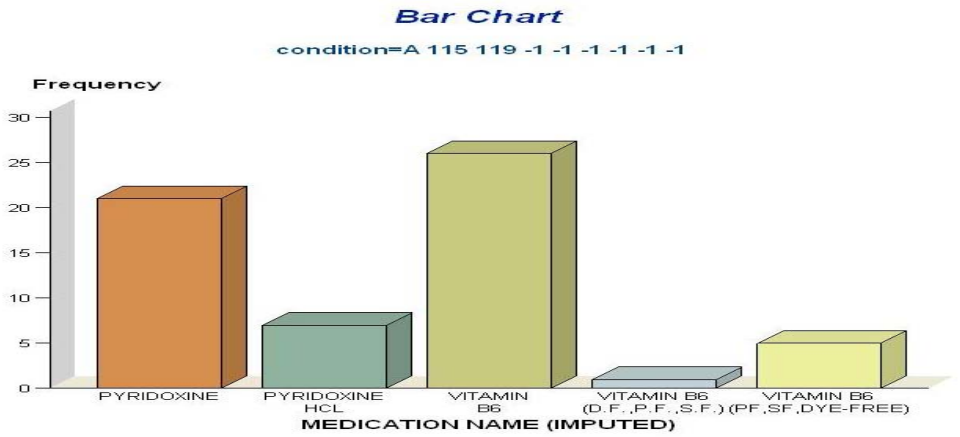


**Figure 2. The medications for Condition=B 1 11 304 -1 -1 -1 -1 -1, the difference between A and B is the risk for pregnancy.**

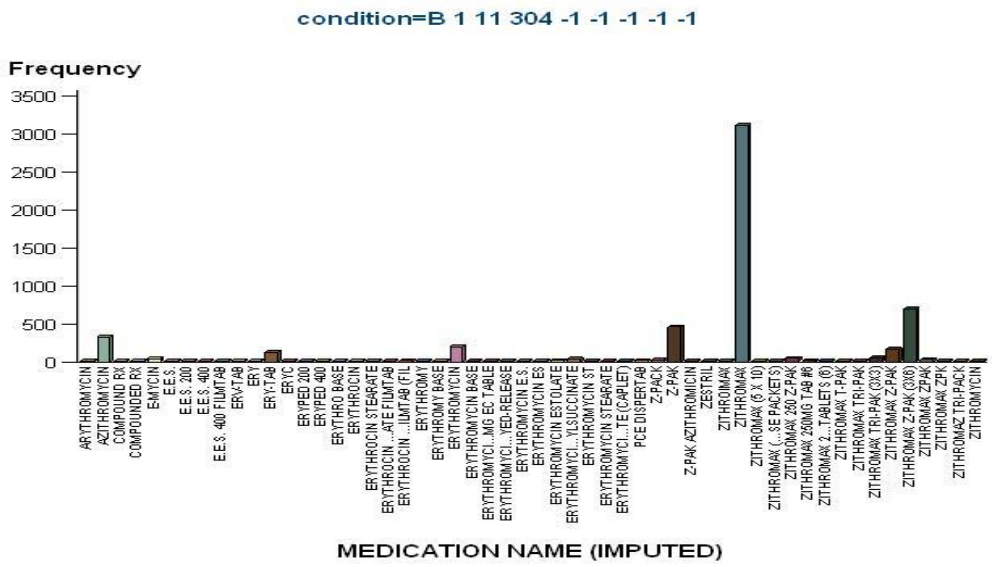
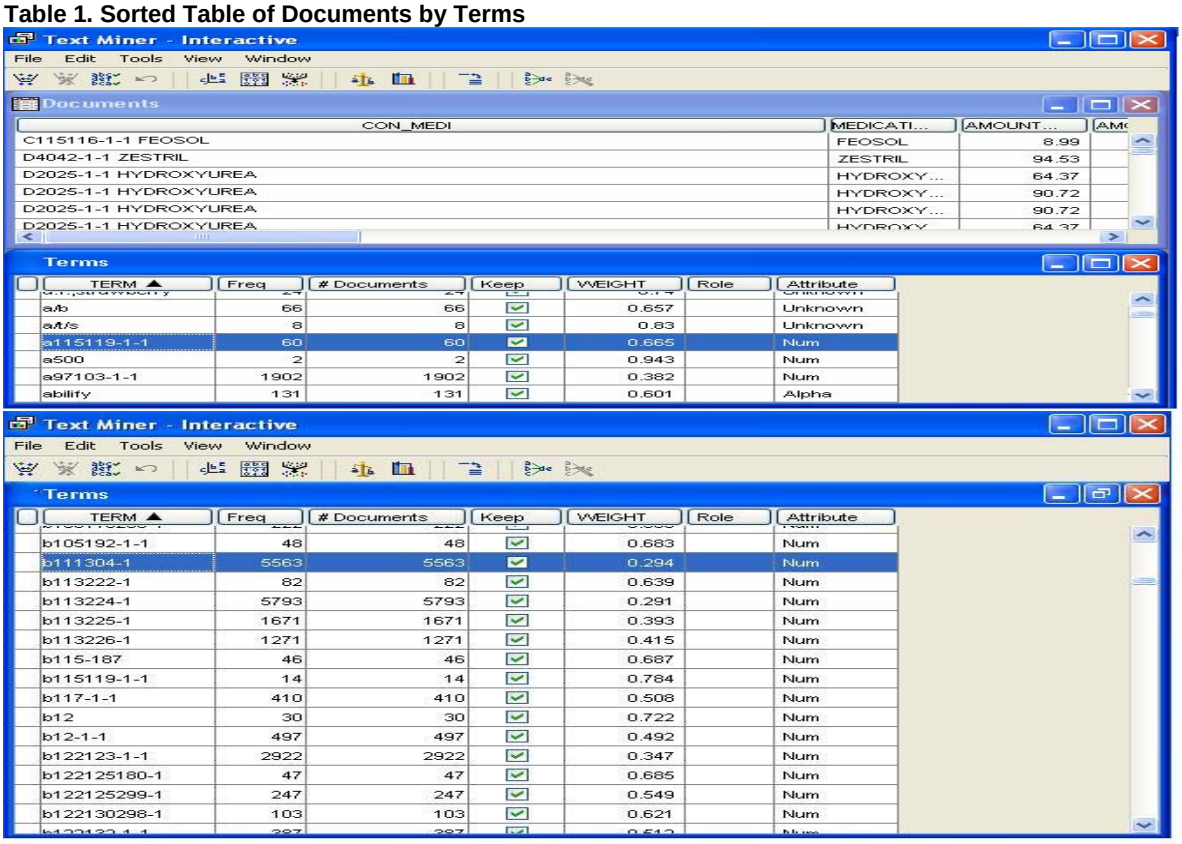



Figure 2 shows the combination of medications for a condition for which there is no evidence of risk in humans for Pregnancy, anti-invectives for therapeutic class #1 , macrolides derivatives for sub-class#1 and macrolides for sub- sub-class#1, is Zithromax , Zithromax(Z.pak(3)(6)) , Z.pak, and Azithromaycin. They are different words for the same antibiotic. It is clear that this is a common antibiotic.

From these two examples, we could find the combination of medications for every condition similarly. In Text Miner, we obtained the related medications with respect to conditions that are similar to the results above.

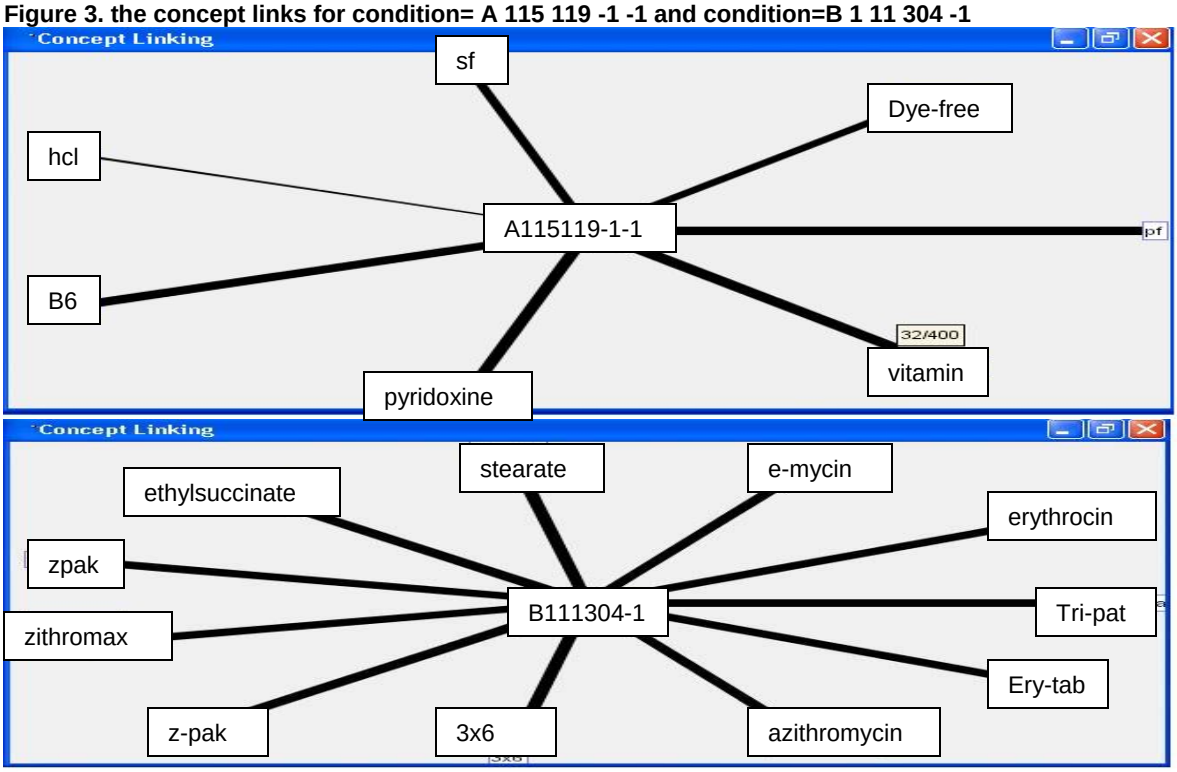
**Table 1. Sorted Table of Documents by Terms**



There are 60 and 5563 observations for condition=A 115 119 -1 -1 and condition=B 1 11 304 -1 in the entire data set respectively.

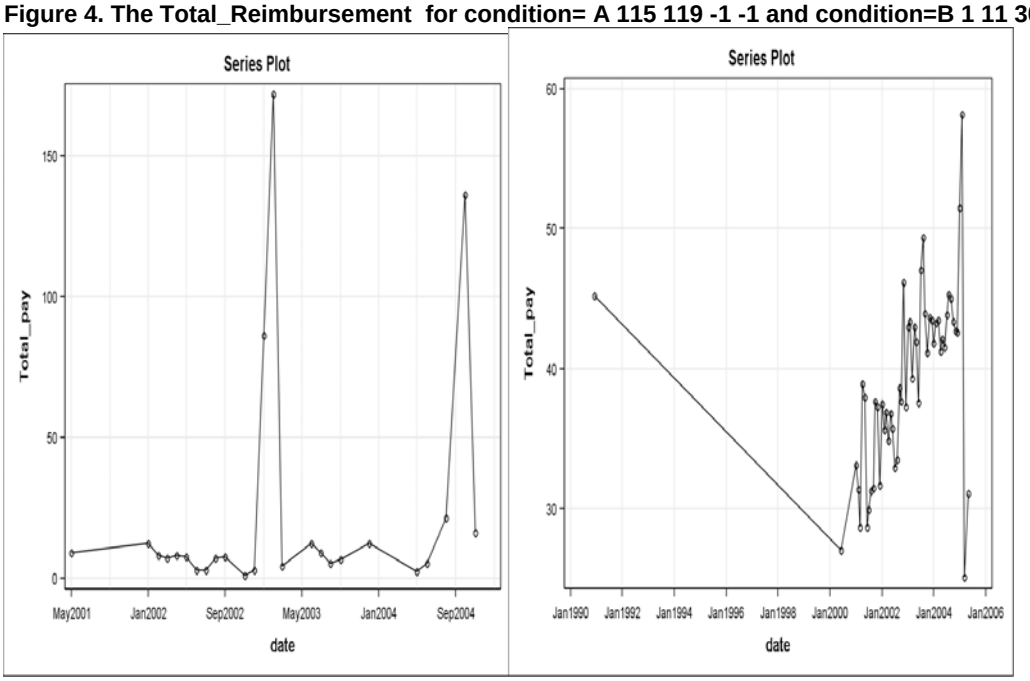
Concept links are graphs consisting of word terms that are connected to each other. The nodes are labeled with descriptive text, representing the "concept". Concept links are an important means of knowledge representation because many people find them intuitive and easy to understand. Concept maps have been used in many fields including education, management, artificial intelligence, knowledge representation, knowledge acquisition, and Linguistics. Text Miner provides the user with the concept links feature. The concept links for condition= A 115 119 -1 -1" and condition=B 1 11 304 -1" are given in Figures 3.

**Figure 3. the concept links for condition= A 115 119 -1 -1 and condition=B 1 11 304 -1**




Note that the medications are Vitamin, Pyridoxine and Zithromax, Zithromax (Z.pak(3)(6)), Z.pak, and Azithromaycin related to these conditions respectively. There were 32 records of Vitamin for condition= A 115 119 -1 -1 and 4166 records of Zithromax for condition=B 1 11 304 -1. That means they are the main medications for these conditions.

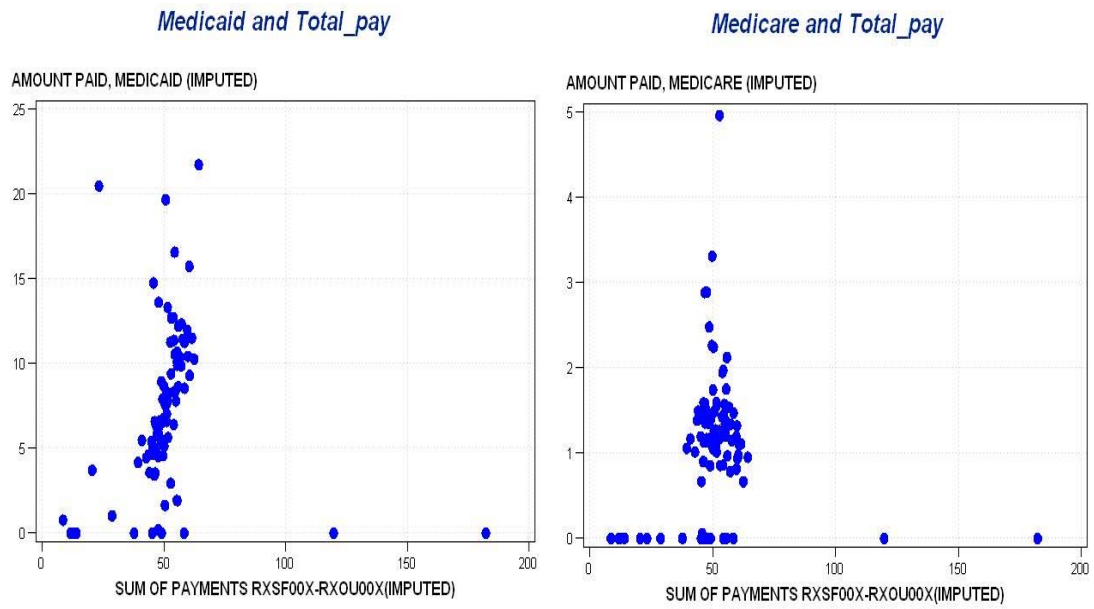
**Figure 4. The Total_Reimbursement for condition= A 115 119 -1 -1 and condition=B 1 11 304 -1**

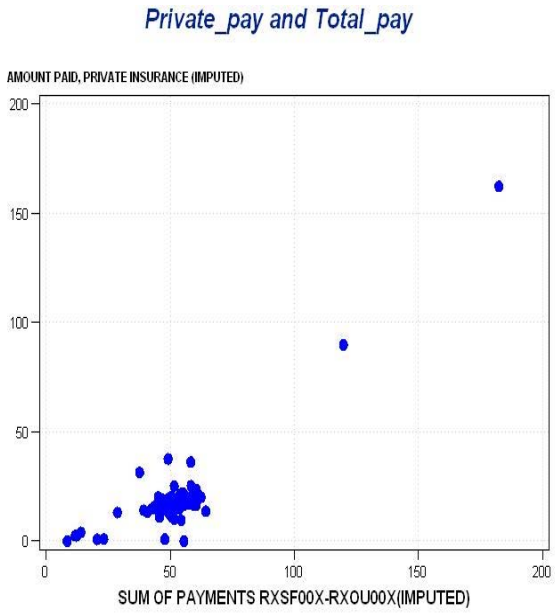


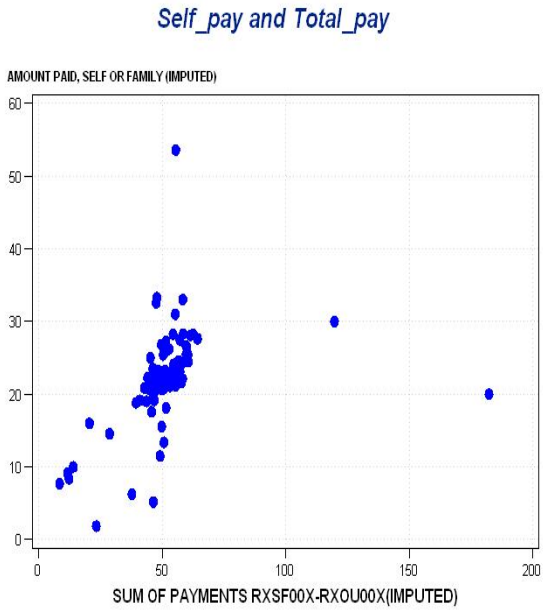

(a) total_pay for condition= A 115 119 -1 -1 (b) total_pay for condition=B 1 11 304 -1

Figure 4 shows that the total pay for Vitamin and Pyridixine keeps stable around $20. However, the total pay for Zithromax, Z-pak, and Azithromaycin increases from 30 dollars to 50.

**Figure 5. The weight of self-pay, Medicaid cost, Medicare cost and private insurance in the total cost.**






Note that the average of Total_pay is around 50 dollars and the average of Self_pay, Medicaid, Medicare and Private Insurance were about 21, 9, 1.4 and 22 dollars respectively. Here the amounts of Self_pay, Medicare and Private Insurance did not change much with respect to date; however, the amount of Medicaid changed considerably according to date from 3 to 15 dollars.

## CONCLUSION

The data set we analyzed includes 317,076 observations from 1996 to 2004. First, we created a new variable based on the data, which included all the information about therapeutics for each observation. After filtering according to this variable, we reduced the dataset to 202,561, in which there is no empty value for that column.

With Enterprise Guide, we sorted the dataset with respect to condition. We found the combination of medications related to the patient therapeutics. For instance, the combination of medications to the condition showed no risk for Pregnancy. Nutritional products were used for therapeutic class #1, Vitamins used for sub-class#1 are PYRIDIXINE and VITAMIN B6.

We can also sort the result in Text Miner by condition and find the exact one we are interested in. Then we can find the combination of medications for every condition through the concept link tool. We have the similar results with the ones we got by using Enterprise Guide.

In the Time Series Analysis, we accumulated using the average. After accumulating, the number of records decreased to 178. Then we used the explore node in SAS Enterprise Miner to explain the weights of the other four parts in the total cost. The weights of Self_pay, Medicare and Private Insurance kept stable and the weight of Medicaid kept increasing according to Total_pay.

## REFERENCE

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, P., *Introduction to use of text miner software.*
3. SAS, *SAS course notes.*

## ACKNOWLEDGMENTS

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

## CONTACT INFORMATION

Guoxin Tang
Ph.D. Student in Applied Mathematics
Department of Mathematics
University of Louisville
Louisville, KY 40292

Office Phone: 502-852-7012
E-mail: g0tang01@louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.