Paper PO12

Text Mining to Summarize Complicated Datasets Containing Structured, Nominal Data

Hamed Zahedi, University of Louisville, Louisville, KY

ABSTRACT

The purpose of this study is to filter a large, healthcare database to a cohort of patients undergoing treatment for Osteomyelitis. There are up to fifteen different columns of nominal data to search for Osteomyelitis.

We used SAS Enterprise Guide and the RXMATCH function to summarize the codes defining Osteomyelitis, using potentially 15 columns of information. An alternative approach is to use SAS Text Miner. We bring all fifteen variables into one column as a string of codes, using the CATX function. Then we use SAS Text Miner on the defined text string; the terms window in the output gives the frequency and number of documents. We use Text Miner features such as “Treating as equivalent terms”, “Sorting” and “Filtering” to get summaries of different diagnosis or procedures.

We started with a total of 117,577 patients. After filtering down to Osteomyelitis (ICD9 code 730) using the strings of diagnoses, there were 44,035 records. By looking at procedure codes, we found 22,649 patients had amputation of lower limb (841). We can also use Text Miner to examine Procedures that have been used the most to treat patients with Osteomyelitis, including 84.1 (Amputation of lower limb) with 22,649 patients and 38.93 (Venous catheterization) with 11,341 patients.

Text Miner significantly reduces the amount of time and coding necessary to extract basic information about patient diagnoses and treatment procedures when contained in multiple data columns and nominal in nature.

INTRODUCTION

Getting a summary of a dataset is usually assumed to be trivial. We usually want to find the frequency of a given input (code) for a variable (column of dataset), or more than one variable in health care data. Using the National Inpatient Sample, the primary diagnosis code is given for each patient, but there are also fifteen other variables to enter other diagnoses for each patient. Working with only one variable is already complicated; considering all fifteen variables is a difficult problem. The purpose of this study is to filter this large, healthcare database to a cohort of patients undergoing treatment for Osteomyelitis. We use SAS Text Miner to demonstrate a simplified method to search these fifteen columns.

METHODS

The data are from the NIS, The Nationwide Inpatient Sample. 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. [1] The data are available for a small charge at http://www.ahrq.gov.

We had five years of data, 2000 to 2004. Each year of data included about 8 million records. We filtered the patients that have Osteomyelitis (bone infection) or amputation using the DRG codes 113, 114, 213, 238 and 285.The Diagnosis-Related Groups (DRG) coding is used in health care data to group different diagnoses. The codes translate to

DRG 113: Amputation for circ system disorders except upper limb & toe
DRG 114: Upper limb & toe amputation for circ system disorders
DRG 213: Amputation for musculoskeletal system & conn tissue disorders
DRG 238: Osteomyelitis
DRG 285: Amputation of lower limb for endocrine, nitrite & metabol disorders

After preprocessing the data, we had 117,577 patient records involving Osteomyelitis with or without amputation. This dataset has 126 variables including fifteen columns of Diagnosis codes and fifteen columns of Procedure codes. Any diagnosis or procedure code can appear in any one of the fifteen columns.

Diagnosis codes are given in a variety of digits/characters. The codes have a 3-digit stem (diagnosis codes) and 2-digits for specifics. Procedure codes have a 2-digit stem. For example, the code “84” represents “Procedures on musculoskeletal system”, 84.1 represents “Amputation of lower limb” and 84.11 is for “Amputation of toe”. The decimal point is omitted in the data columns in the dataset. There are also some codes starting with the letter “V” or “E”. It is not possible to treat these codes as numbers, and they have to be considered as strings of characters. Working with these types of datasets requires a large amount of time for preprocessing.

We used SAS Enterprise Guide and COUNT, RXMATCH and SUBSTRN, along with other functions in several lines of code to get a summary (count) of the codes defining Osteomyelitis, using only the primary diagnosis.

> COUNT: Counts the number of times that a specific substring of characters appears within a character string specified. [2]
> RXMATCH: Finds the beginning of a substring that matches a pattern. [2]
> SUBSTRN: Returns a substring, allowing a result with a length of zero. [2]

There are several lines of code needed to use these functions. We have to use caution since there are possibilities of double counting or of miscounting some codes using these functions. For example; in using the COUNT function, if two occurrences of the specified substring overlap in the string then inconsistent results will be returned. [2] For example, for finding the number of substrings, “88”, in the string “888”, COUNT ('888', '88') might return either 1 or 2.

Another example occurs in using RXMATCH. If we look for ‘730’ in codes, this function counts the code 2730 too, but we want only codes that start with 730 and we do not want 2730 included. This method can be performed on data using only the primary diagnosis, but as we mentioned, there are fifteen of these diagnosis coded to be considered in our dataset. Without any concatenation of the fifteen columns, we have to repeat the procedure for each column of Diagnosis codes and again for each column of Procedure codes to get the summary.

The best way to work with all fifteen 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. [2] The expression below shows an example of using this function to create a text string for the procedure codes.

> CATX (‘ ’ , finalhcup.PR1 , finalhcup.PR2 , finalhcup.PR3 , finalhcup.PR4 , finalhcup.PR5 , finalhcup.PR6 , finalhcup.PR7 , finalhcup.PR8 , finalhcup.PR9 , finalhcup.PR10 , finalhcup.PR11 , finalhcup.PR12 , finalhcup.PR13 , finalhcup.PR14 , finalhcup.PR15)

This expression makes a string of codes from Column PR1 to PR15 with a space between them. We also do the same for the fifteen diagnosis variables. As a result, we have two new columns carrying strings of procedures and diagnosis.

We were interested in Osteomyelitis only, so we needed to keep the records that have Osteomyelitis code (730) in any of the fifteen columns of diagnosis, so we filtered the created text strings where a defined diagnosis substring containing 730, using the code below

```
data sasuser.Osteomyelitis;
set sasuser.Query_For_finalhcup;
if (rxmatch('730',STR_DX) > 0) then osteom=1;
else osteom=0;
run;
```

We created a new column that indicates whether a patient is diagnosed with Osteomyelitis or not. We may count the number of patients with Osteomyelitis by filtering the data on the defined column as 1. In the filtered dataset, we have 44,035 records containing at least one diagnosis for Osteomyelitis. To consider the patients with amputation of lower limb (841), we should use the similar code for the string of procedures and create a column to indicate whether patients had amputation or not, using the code below

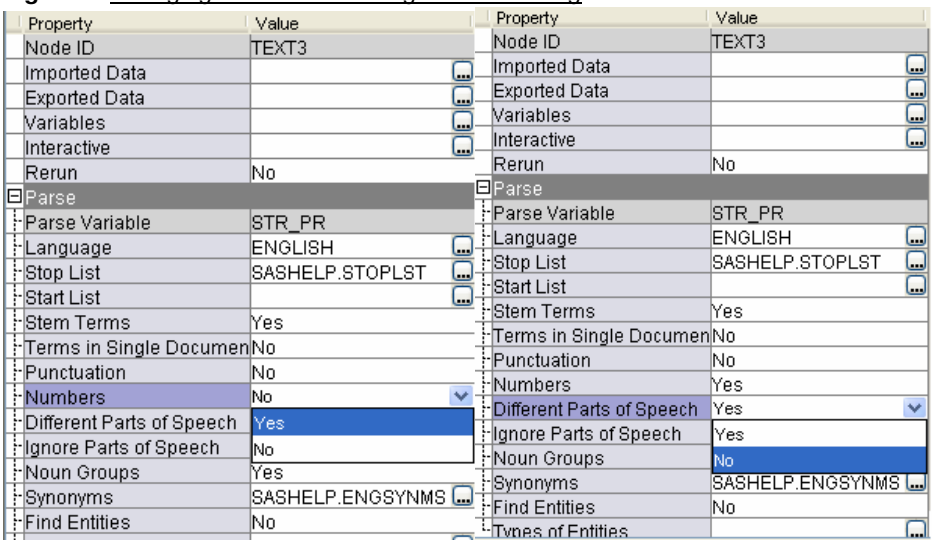
```
data sasuser.amputation;
set sasuser.Query_For_Osteomyelitis;
if (rxmatch('841',STR_PR) > 0) then amp=1;
else amp=0;
run;
```

Then we count the number of records that have a one to represent amputation. There were a total of 22,662 patients with amputation of lower limb (841). To investigate other diagnosis and/or other procedures, we need to repeat the above procedure to get the result, and it can be time consuming.

RESULTS

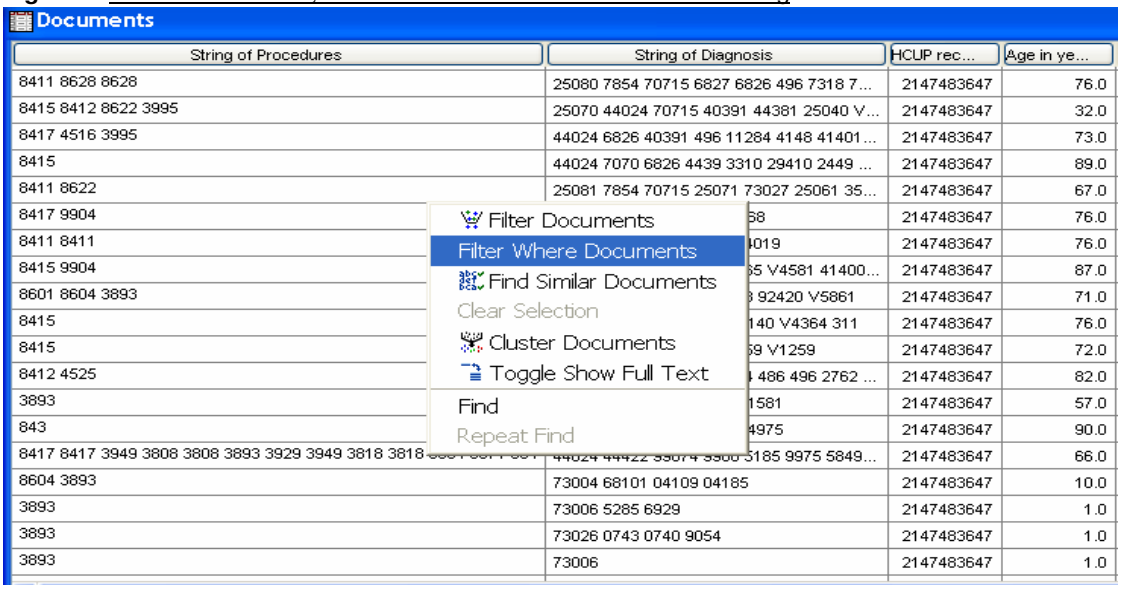
An alternative approach is to use SAS Text Miner. We use the dataset, including the concatenated columns defined using the CATX function, in Enterprise Miner. Then we use SAS Text Miner on the defined text string; since we work with numbers (ICD9 codes), we had to change the default setting on Text Miner and switch Numbers from No to Yes. We also had to change "Different Parts of Speech" from Yes to No, so we would be able to put the codes with or without the letters "V" and "E" in an equivalent group. Figure 1 shows the changes to the default settings for Text Miner.

Figure 1. Changing the Default Setting for Text Mining



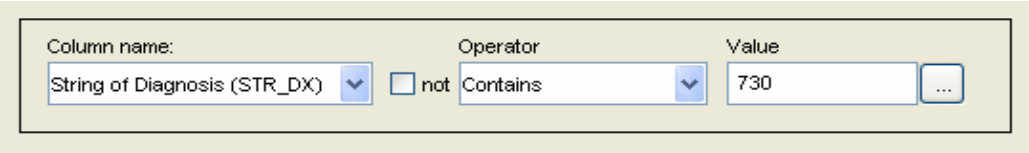
The terms window in the output gives the frequency and number of documents for each code. We were interested in Osteomyelitis only, so we needed to keep the records that have an Osteomyelitis code (730) in any of the fifteen columns of diagnosis. Thus, we filtered in Text Miner where the defined diagnosis substring contained 730. We did this by right clicking on the document window and using the option of "Filter Where Documents" as shown in figure 2.

Figure 2. Document Window; Select "Filter Where Documents" for filtering



We then choose the strings of diagnosis codes containing 730 to filter the selected rows from the documents as shown in figure 3.

Figure 3. Filtering for Osteomyelitis (ICD9 code 730)

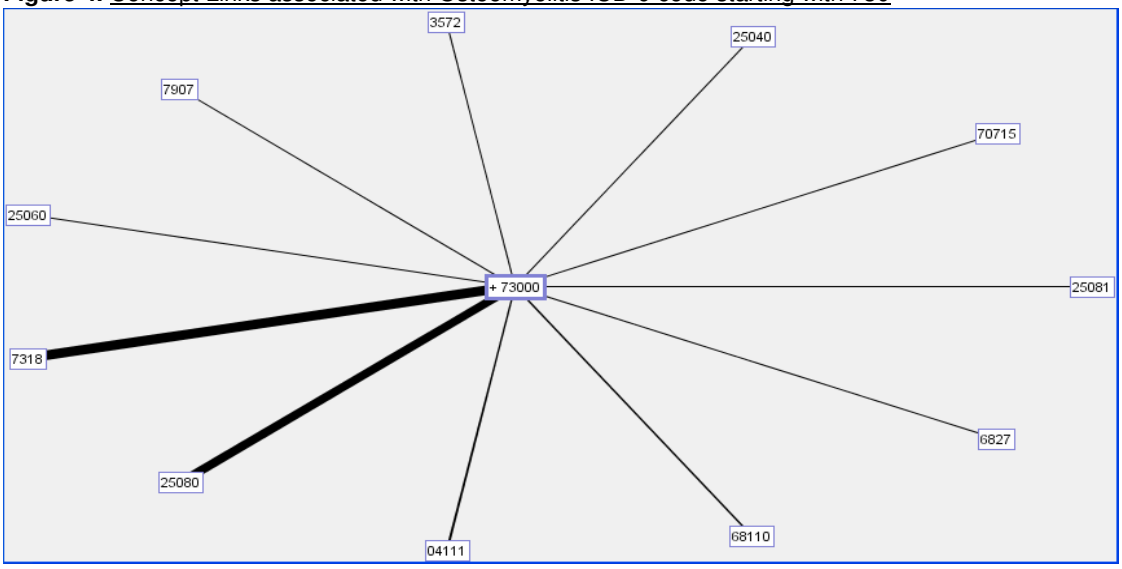


After filtering, we had 44,035 patients with Osteomyelitis in this dataset. Osteomyelitis (730) might have been recorded in any of fifteen columns of diagnosis codes. Enterprise Miner let us study the relationship between the terms as concept links.

We can view terms that are highly associated with the selected term in a hyperbolic tree display. The tree display shows the selected term in the center of the tree structure. The selected term is surrounded by the terms that correlate the strongest with it. Using the tree display, we can select a term associated with the first term to view its associated terms, and so on.

We consider all the codes starting with 730, and study the concept links of them. Figure 4 shows the concept links associated with ICD-9 code, 730. Among those terms, we see many of them start with 250. This code represents "Diabetes mellitus". In particular, the link to 25080 shows a bigger portion of relationships compared to other links; ICD-9code 25080 represents "Diabetes with other specified manifestations". Another "bold" link is 7318, which is associated with "Other bone involvement in diseases classified elsewhere".

Figure 4. Concept Links associated with Osteomyelitis ICD-9 code starting with 730



Diagnosis codes have a 3-digit stem and 2-digits for detail, so the osteomyelitis code appears in many different codes, but they all start with 730. Table 1 shows the list of these different codes in our dataset. The most frequent code is 73027. We study this code by looking at concept links associated to it, shown in Figure 5. We see different terms, and three of them start with 250. We can expand the links to see relationships between these codes with others, shown in Figure 6.

Figure 5. Concept Links associated with code 73027

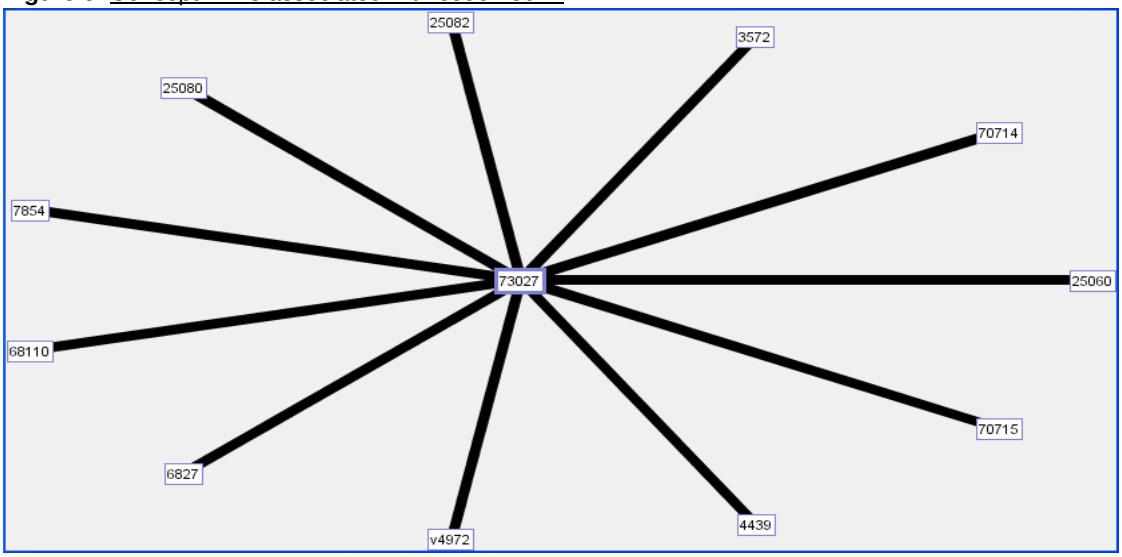
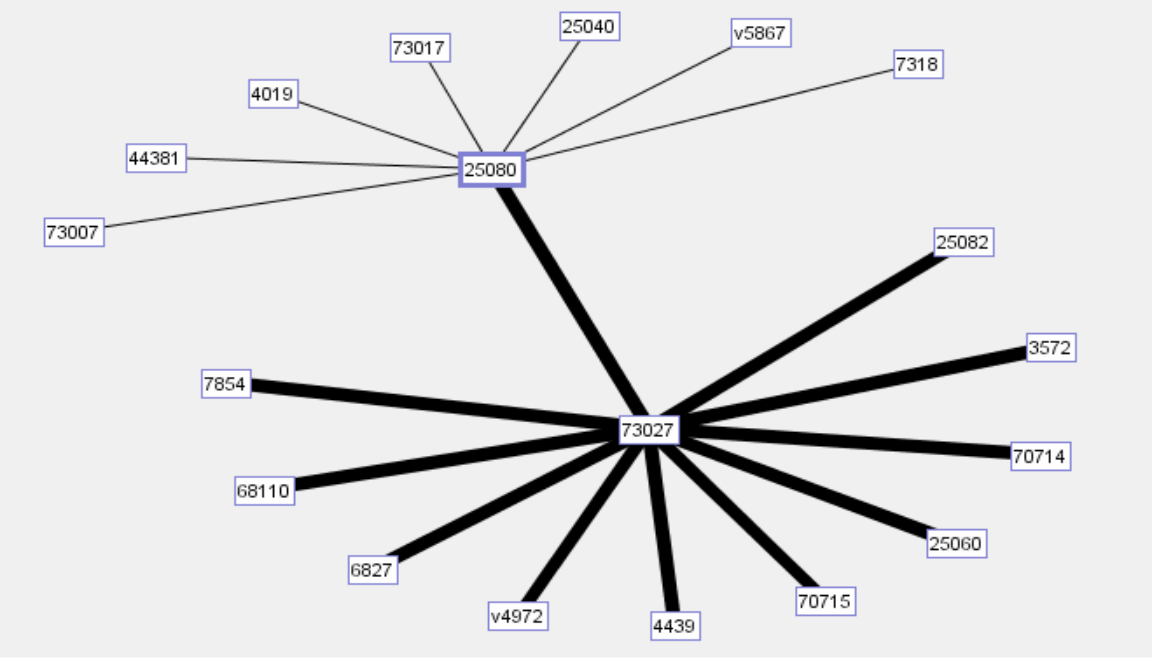


Table 1. List of Osteomyelitis codes appear in the dataset

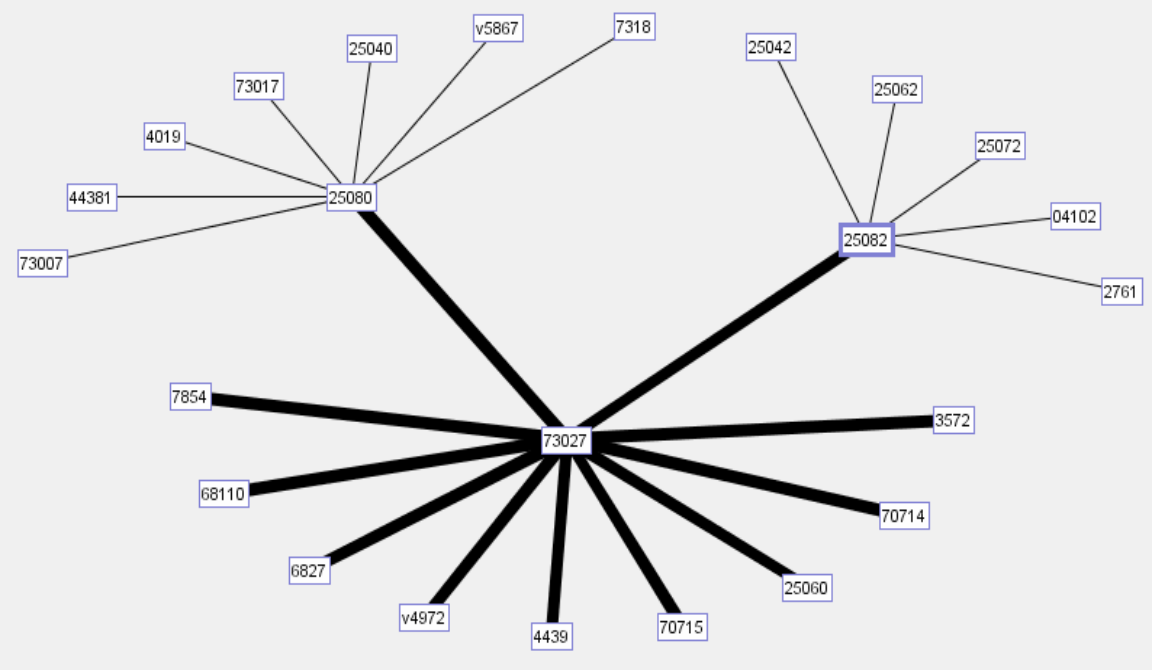
TERM ▲	Freq	# Documents	Keep	WEIGHT	Role	Attribute
73000	23	23	☑	0.707		Num
73001	85	85	☑	0.585		Num
73002	67	67	☑	0.607		Num
73003	60	60	☑	0.617		Num
73004	332	332	☑	0.457		Num
73005	505	505	☑	0.418		Num
73006	1301	1301	☑	0.329		Num
73007	9671	9669	☑	0.142		Num
73008	1047	1047	☑	0.35		Num
73009	37	37	☑	0.662		Num
73010	30	30	☑	0.682		Num
73011	42	42	☑	0.65		Num
73012	69	69	☑	0.604		Num
73013	30	30	☑	0.682		Num
73014	97	97	☑	0.572		Num
73015	696	695	☑	0.388		Num
73016	1411	1411	☑	0.322		Num
73017	4999	4998	☑	0.204		Num
73018	508	508	☑	0.417		Num
73019	38	38	☑	0.66		Num
73020	159	159	☑	0.526		Num
73021	158	158	☑	0.527		Num
73022	195	195	☑	0.507		Num
73023	172	172	☑	0.519		Num
73024	627	627	☑	0.398		Num
73025	1470	1468	☑	0.318		Num
73026	2502	2502	☑	0.268		Num
73027	16260	16258	☑	0.093		Num
73028	2610	2609	☑	0.264		Num
73029	69	69	☑	0.604		Num
73036	2	2	☑	0.935		Num
73037	10	10	☑	0.785		Num
73080	3	3	☑	0.897		Num
73082	2	2	☑	0.935		Num
73083	2	2	☑	0.935		Num
73087	31	31	☑	0.679		Num
73088	143	143	☑	0.536		Num
73089	2	2	☑	0.935		Num
73090	2	2	☑	0.935		Num
73095	8	8	☑	0.806		Num
73096	13	13	☑	0.76		Num
73097	18	18	☑	0.73		Num
73098	8	8	☑	0.806		Num
73099	2	2	☑	0.935		Num

Figure 6. Concept Links associated with 73027 and 25080



We can also expand links for 25082, as it is shown in Figure 7. We see other codes starting with 730 and 250 associated with this code.

Figure 7. Concept Links associated with 73027, 25080 and 25082

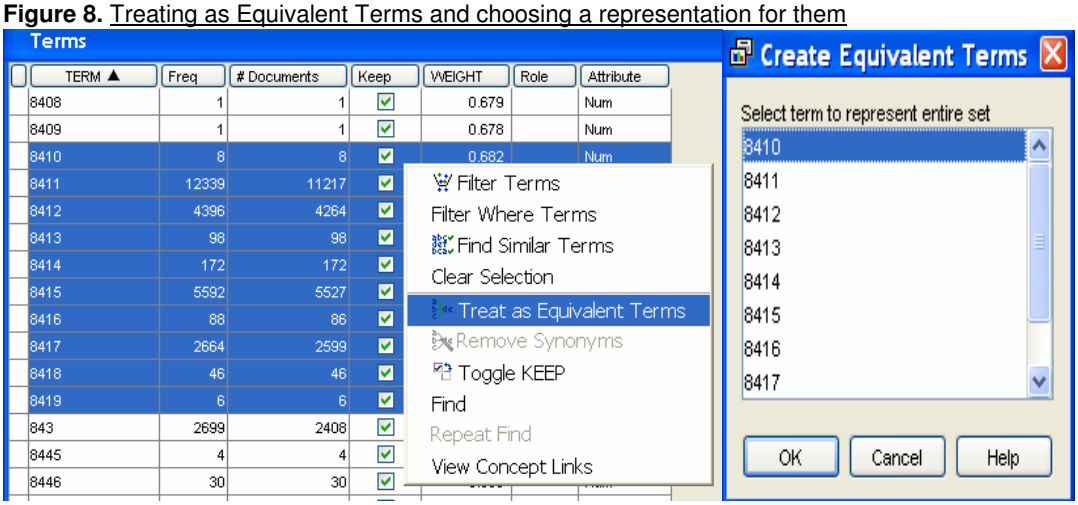


This procedure can be done to study the relationship between any term with other terms.

After filtering, we have the dataset that contains patients with Osteomyelitis, and we need to investigate the procedures that have been done on these patients, especially amputation in the lower limb (841). In this case, the codes we have to consider are 841, 8410, 8411, 8412 ... 8419. The fourth digit indicates the location of the amputation in the lower limb.

We sort the results by Terms to find these codes. We have the number of patients for each of these codes, but there are some patients who are having more than one of these procedures, so if we add all the numbers of these rows, we might count some patients twice or more. To avoid this, we treat these terms as equivalent by selecting them and then right clicking on terms and choosing “Treat as Equivalent Terms” as shown in figure 8. We choose 8410 to represents this group.

Figure 8. Treating as Equivalent Terms and choosing a representation for them



Out of 44,035 patients with Osteomyelitis, 22,649 patients had an Amputation of the lower limb; over 51 % of the patients with Osteomyelitis.

To see what procedures have been done on these patients, we sort by number of documents in the Terms window. We have the group of 84.1 (amputation of lower limb) with 22,649 patients and 38.93 (Venous catheterization) with 11,341 patients in the first two places followed by 9904 (Transfusion of packed cells) with 4,302 patients, 8622 (Excisional debridement of wound, infection, or burn) with 3,907 patients, 3995 (Hemodialysis) with 2,787 patients and 843 (Revision of amputation stump) with 2408 patients. The sorted results are shown in Table2.

Table 2. Sorted Table of Terms by Number of Documents

TERM	Freq	# Documents ▼	Keep	WEIGHT	Role	Attribute
8410	25409	22649	☑	0.682		Num
3893	11832	11341	☑	0.158		Num
9904	4442	4302	☑	0.166		Num
8622	4756	3907	☑	0.205		Num
3995	3146	2787	☑	0.194		Num
843	2699	2408	☑	0.199		Num
8604	2198	2057	☑	0.3		Num
9921	1869	1836	☑	0.311		Num
8848	1887	1787	☑	0.217		Num
8628	1232	1116	☑	0.346		Num
8842	1039	1030	☑	0.265		Num
9214	1039	1028	☑	0.39		Num

We also investigated the previous amputations as identified in the diagnosis codes; we took the resulting dataset from Text Miner and did the text mining again, but this time on the strings of diagnosis to find the number of patients with previous amputations. From the total of 44,035 patients, 4,673 patients had previous amputation. We used ICD9 codes of 895, 905.9, and 997.6, E878.5, V49.6 and V49.7 to find previous amputations. The ICD9 codes translate to:

895 Traumatic amputation of toe(s) (complete) (partial)
905.9 Late effect of traumatic amputation
997.6 Amputation stump complication
E878.5 Amputation of limb(s)
V49.6 Upper limb amputation status
V49.7 Lower limb amputation status

Table 3 shows the terms for the above ICD9 codes.

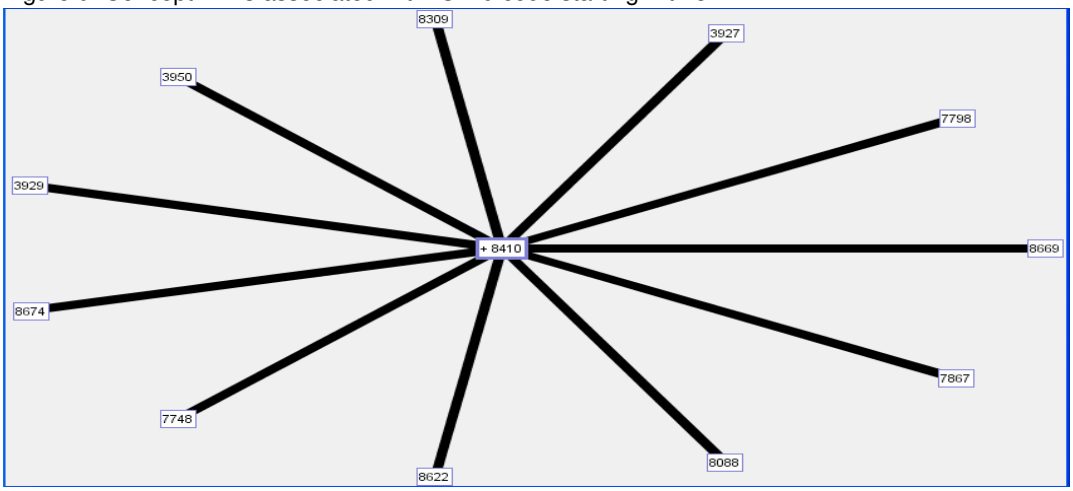
Table 3. Previous Amputations of patients with Osteomyelitis

TERM	Freq	# Documents	Keep	WEIGHT	Role	Attribute
v4970	5745	4673	☑	0.711		Num
v4966	3	3				
v4960	3	3				
99760	16	16				
e8785	539	505				
v4973	197	197				
v4965	5	5				
v4961	5	5				
v4970	22	22				
v4974	5	5				
99769	691	691				
v4977	6	6				
99762	1742	1742				
9059	8	8				
v4972	837	837				
99761	9	9				
v4975	896	896				
v4971	338	338				
v4976	368	368				
8951	3	3				
v4962	49	49				
v4963	3	3				

Now we view terms that are highly associated with the selected term in a hyperbolic tree display. We consider all the codes starting with 841 and study the concept links of them.

Figure 9 shows the concept links associated with the ICD-9 code, 841. Among those terms, we see many of them start with 77, 39 and 86. These codes represent “Incision, excision, and division of other bones”, “Other operations on vessels” and “Operations on skin and subcutaneous tissue” respectively. These are operations that have been performed on patients with amputation.

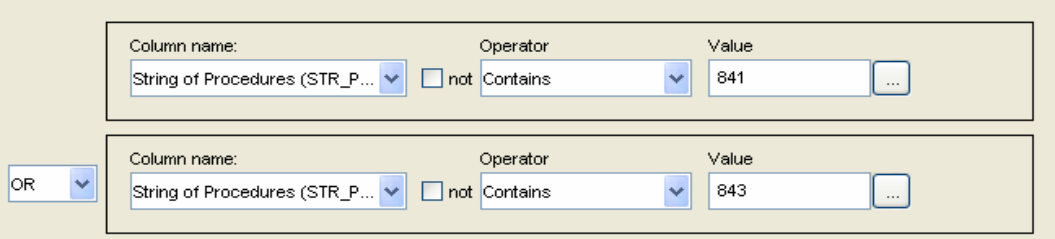
Figure 9. Concept Links associated with ICD-9 code starting with 841



We consider previous amputations for the patient with current amputation, so we filtered the procedure codes with ICD9 Codes of 84.1 and 84.3 as shown in figure 10.

84.1 Amputation of lower limb
84.3 Revision of amputation stump

Figure 10. Filtering for Amputation (ICD9 codes 841 or 843)



We have a total of 23,773 patients. There are 3,586 patients with previous amputation (ICD9 Codes of 895, 905.9, and 997.6, E878.5, V49.6 and V49.7). Table 4 shows the terms table of the result for these patients.

Table 4. Previous Amputations of patients with Current Amputation

TERM ▲	Freq	# Documents	Keep	WEIGHT	Role	Attribute
v4970	4511	3586	☑	0.711		Num
v4971	226	226				
99762	1595	1595				
e8785	506	474				
v4972	554	554				
v4975	599	599				
99769	646	646				
v4970	7	7				
99761	8	8				
9059	6	6				
v4966	1	1				
v4973	130	130				
v4976	190	190				
v4974	2	2				
v4965	2	2				
v4962	19	19				
99760	12	12				
v4961	3	3				
v4963	2	2				
8951	3	3				

There are some effective antibiotics to treat Osteomyelitis. Our result shows Amputation assumed to be the primary treatment of Osteomyelitis. Of the 44,035 patients, 22,649 patients had the amputation of lower limb and 1,836 patients had the injection of antibiotics (ICD9 code 99.21). Only 10 patients had Injection or infusion of the Oxazolidinone class of antibiotics (Linezolid injection) with ICD9 code 00.14.

Comparing the number of patients with amputation and the number of patients with effective antibiotics injections, we observed that in many cases, amputation is performed without trying the different antibiotics for treatment of Osteomyelitis before amputation.

Some antibiotics must be used in hospitals and by having the supervision of experts, for example, Vancomycin, and some antibiotics can be used orally at home, for example, Linezolid (Zyvox). Choosing the right antibiotics and the right length of treatment are very important parts of the treatment. Making the right choice of antibiotics will decrease the number of amputations in the treatment of Osteomyelitis.

CONCLUSION

We started with 117,577 records (patients). After filtering down to Osteomyelitis using the strings of diagnosis codes, we reduced the dataset to 44,035 records. By looking at procedure codes, we found 22,649 cases have had amputation, which is more than half (51%) of the patients with Osteomyelitis.

We can also sort the result in Text Miner by the number of documents and see that the first two procedures that were performed on patients with Osteomyelitis are 84.1 (Amputation of lower limb) with 22,649 patients and 38.93 (Venous catheterization) with 11,341 patients. We found only 3,133 patients with an injection of antibiotics and only 10 patients with Linezolid (Zyvox) injection.

Text Mining in Enterprise Miner is a great tool to get data summaries from similar data sets and does not require the programmer to write long codes. We use Text Miner features such as “Treating as equivalent terms”, “Sorting” and “Filtering” to get summaries of different diagnosis or procedures.

Considering the result on this particular dataset, amputation was performed the most frequently compared to other procedures for patients with Osteomyelitis. Physicians assume amputation is the first (and best) treatment for Osteomyelitis. Injection of antibiotics was performed on only about 4% of the patients with Osteomyelitis. In many cases, infection has recurred and amputation was performed more than once.

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. SAS Enterprise Guide Help Menu

3. ICD9.chrisendres.com

ACKNOWLEDGMENTS

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

CONTACT INFORMATION

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

Hamed Zahedi
Ph.D. Student in Applied Mathematics
Department of Mathematics
University of Louisville
Louisville, KY 40292
Office Phone: 502-852-3519
E-mail: hamed.zahedi@louisville.edu