Paper 001-29

Solutions to the Investigation of Healthcare Outcomes in Relationship to Healthcare Practice

Patricia B. Cerrito, University of Louisville, Jewish Hospital Center for Advanced Medicine, Louisville, KY

ABSTRACT

Data mining can be used to examine the relationship between physician practice and patient outcomes while accounting for variability in patient risk and examining variability in physician practice. For example, changes in physician prescribing patterns are easily visualized using data mining tools. Associations can be used to target interactions and/or outcomes for additional examinations. Patterns of prescribing can be tracked using visualization techniques. Once the prescribing patterns are known educational materials can be developed to support necessary changes in the prescribing patterns, change buying habits to reduce the cost of goods, and provide the patient with the most effective, cost saving medications.

Text mining can be used to examine unstructured notes that are often contained within patient charts, and to examine diagnosis codes that are used for billing purposes. Another use of the text mining is to compare different sources of medical information to determine optimal sources to examine so that physicians (and consumers) can keep current with practice guidelines. Various data mining tools will be illustrated with examples from medical data. In particular, the relationship between practice and outcome will be examined, so that best practices can be identified that will result in the most optimal outcomes.

INTRODUCTION

So much medical information is published that it is difficult for physicians to keep current. Automatic processing of relevant information from the medical literature will enable physicians to find quickly what they need. Keyword searches of Medline, the database maintained by the National Library of Medicine can return thousands of abstracts for review. Text mining can refine the search to a handful of documents that need to be scanned.

Drug formularies often make newer drugs unavailable to patients, particularly those patients who are routinely under-served by the healthcare industry. HMOs originally tried to cut healthcare costs by using tight formularies with few drugs and no allowance for patients who needed drugs outside of the formulary. Recently, pharmaceutical companies have been marketing some of their newer drugs directly to consumers, possibly increasing the friction between patients, physicians, and insurance providers. Once on the formulary, use of the drug is not restricted, and changes to the formulary occur slowly. Usually, decisions are made by administrative committees, sometimes referring to pharmacological information.[1] Health outcomes are rarely considered. Insurance providers often require patients to start on a formulary drug, and to exhibit a high level of adverse events before providing for a patient to go off formulary for a newer drug that reduces the adverse events. It is the purpose of this project to develop a software product that can use clinical data to examine health outcomes in relationship to drugs proposed for addition to the formulary.

A second major problem deals with improper prescribing by physicians. One study found that as many as 81 out of 100 patients improperly received fluoroquinolone with no indication.[2] Continued examination will make a formulary more efficient by monitoring proper usage.[3-5] It will also allow formularies to be changed as needed by discontinuing medications that have little impact on patient outcomes. The formulary can be used to discourage inappropriate prescribing by limiting the uses of individual medications.[6, 7]

Prescribing practices can be related to the severity of patient risk factors. Patients with more severe co-morbidities will require more medications. However, there are many different combinations of medications, and there is substantial variability in how physicians treat individual patients.[8] By an examination of the variability and its relationship to patient outcomes, best practices can be defined while simultaneously reducing costs. Prescribing practices related to poor patient outcomes can be discontinued.[9-13] There are a variety of methods to examine relationships. This paper will demonstrate how data mining techniques can be used effectively to increase healthcare intelligence and to improve the quality of patient care.

Another application of data and text mining is to the examination of healthcare quality. Currently, standard statistical measures such as regression are used to compare healthcare providers. However, such statistical measures assume that the data are collected from a random sample. That assumption requires the further assumption that all providers enter patient information uniformly. This condition is almost impossible to satisfy.

Text Miner to Investigate the Medical Literature

It is difficult for a clinician to read all of the relevant literature that pertains to patient care; there are far too many papers published on a regular basis. Text Miner can help to sort through Medline to find the most important. In this section, a brief introduction will be given to the use of Text Miner, and then an example of a Medline Search will be demonstrated. The initial icons for text miner are given in Figure 1.

Figure 1. Text Miner Node

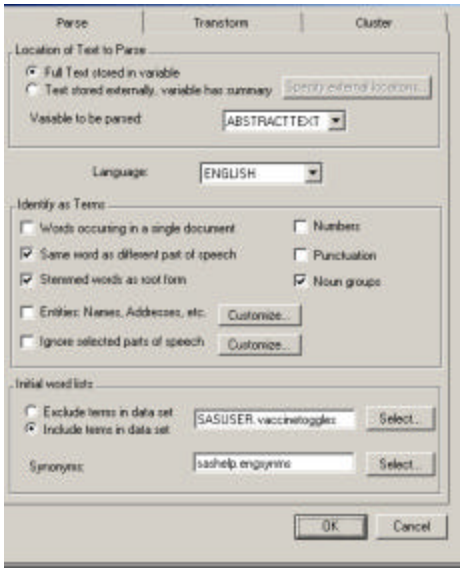


Figure 2. First Settings Screen

The Text Miner Node has three settings screens to examine (Figures 2-4). There are a number of defaults to consider. A standard "stoplist" dataset will remove common words such as "and" and "the" from consideration. The user can add words to the stoplist as needed. A second default is to exclude consideration of words that only occur in one document since those words cannot be used to group documents together. Numbers and punctuation are not ordinarily used to cluster text documents as well.

There is an option to choose if the text is stored in a SAS dataset, or if there is a variable in the dataset that points to the location of the document. This second option is available to reduce the required storage size for a SAS dataset. In the second option, there is no limit on the size of each document; for the first option the size is restricted to 10 pages.

It is also possible to restrict attention to some specific terms by listing them in a dataset.

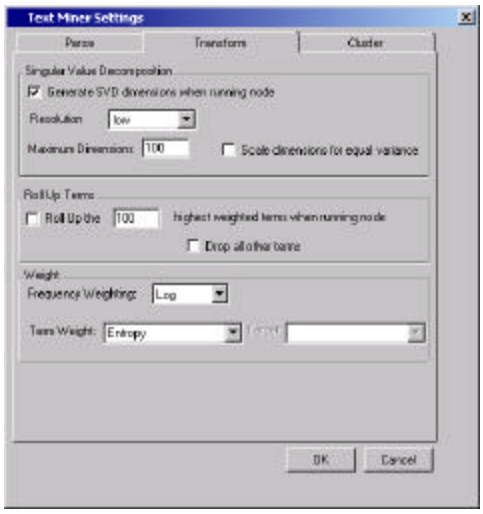


FIGURE 3. SECOND SETTINGS SCREEN

The second screen allows for the user to determine the method of reducing the wordlist matrix to a manageable size. The default is to use singular value decomposition. There are also several possible methods to weight the value of each term in the documents.

To investigate how these weights and methods impact outcomes, it is best to use one dataset and change the settings to see how the results differ.

The number of dimensions defaults to 100. However, that number can be increased for a smaller number of documents, and increased for a large number (although the time factor will increase considerably.

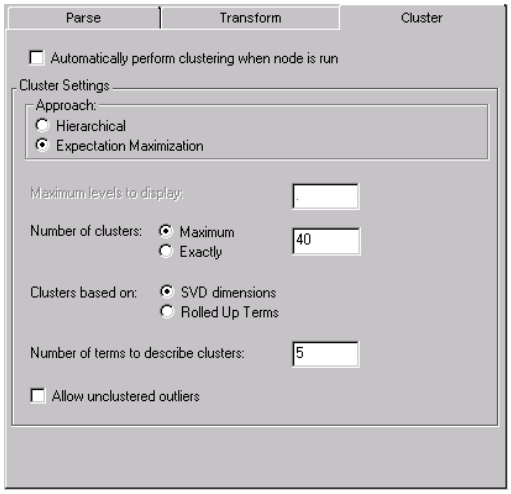


FIGURE 4. THIRD SETTINGS SCREEN

Unless the box is checked, clustering is not automatically performed. However, once Text Miner completes the parsing and transformation steps, the user can request that the clustering be performed.

The user can also set the number of clusters, and the method on which to base the clusters. The default number of terms used to describe the clusters is set at 5. That number may be too small to be able to label the clusters effectively, and it is recommended that this default be increased to 20 or more terms.

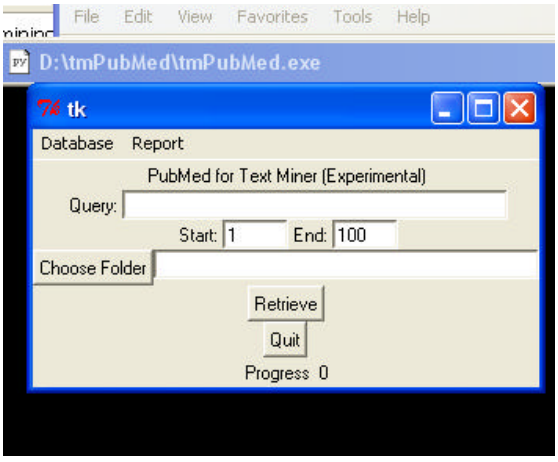
Again, the user is encouraged to work with the defaults to determine their impact upon the results.

The next step is to find a suitable dataset for text analysis. The macro, %tmfilter has been developed to "crawl" through the internet to find documents related to an initial internet site. If an abstract database uses URLs instead of frames,%tmfilter will not work and modifications have to be made. It is also possible to investigate abstracts identified by specific keywords. If the documents are already in some electronic format, including Microsoft Word, %tmfilter can also create a dataset provided that the necessary documents are stored in a folder that can be accessed. Three examples are provided here. The first example was an examination of internet sites that were related to the flu vaccine using the %tmfilter statement to search starting with the URL : http://directory.google.com/Top/Health/Conditions_and_Diseases/Infectious_Diseases/Viral/Influenza/?tc=1 which was available as of January, 2004.

```
%tmfilter(url=http://directory.google.com/Top/Health/Conditions_and_Diseases/Infectious_Diseases/Viral/Infl
uenza/?tc=1,
depth=2,dir=c:\vaccine\dir,
destdir=c:\vaccine\destdir,norestrict=1,
dataset=work.vaccinewebcrawl);
```

Note the option of depth=2. That restricts the search to links of links. The crawl returned 1100 documents. If the depth is increased, so is the time required to crawl, and the amount of storage space. If the depth is expanded to 4, over 50,000 documents were returned and stored in the folder c:\vaccine\dir. Also, by the time the link gets to the depth of 4, the relationship to the initial term of "flu vaccine" becomes very weak. Therefore, such a depth is rarely necessary. Once %tmfilter has completed its run, the text miner node can be used, making sure that the box "text stored externally…" is checked.

Unfortunately, because PubMed (www.PubMed.gov) uses frames, %tmfilter will not work directly with Medline. Russell Albright of the SAS Institute developed a script that can be used to retrieve files from PubMed (Figure 5).



Although the program defaults to a limit of 100 documents, it will retrieve as many as desired. The user can choose a folder in which to store the returned abstracts. A search was conducted on the keywords "treatment+MRSA". MRSA is a bacterial infection that typically occurs in a hospital setting (although not always). It is highly resistant to most antibiotics and is extremely expensive to treat. A search returned almost 1800 abstracts. The statements

```
filename fetch 'd:\MRSA\fetch.xml';
filename SXLEMAP 'd:\tmpubmed\pubmed.map';
libname   fetch xml xmlmap=SXLEMAP
access=READONLY;
run;
```

were used to place the abstracts into a SAS dataset in the library fetch under the name Pubmedarticle. The dataset contained a total of 1737 records. The text abstracts were analyzed in Enterprise Miner using the Text Miner node.

Expectation Maximization was used to cluster the documents, and the default Entropy weights were used for analysis. A total of 1751 terms were listed in the output window (Figure 6).

There are three separate windows that are displayed. The top window displays the complete dataset. The window in the bottom left displays all terms that occur in at least two documents. The number of times the term appears together with the number of documents it appears in is also given. The "keep" field indicates whether the term was used to cluster the documents (designated by "Y"). The bottom right screen gives the clusters identified in the analysis. The problem is to determine what the terms mean.

Figure 6. Output Window for Text Miner.

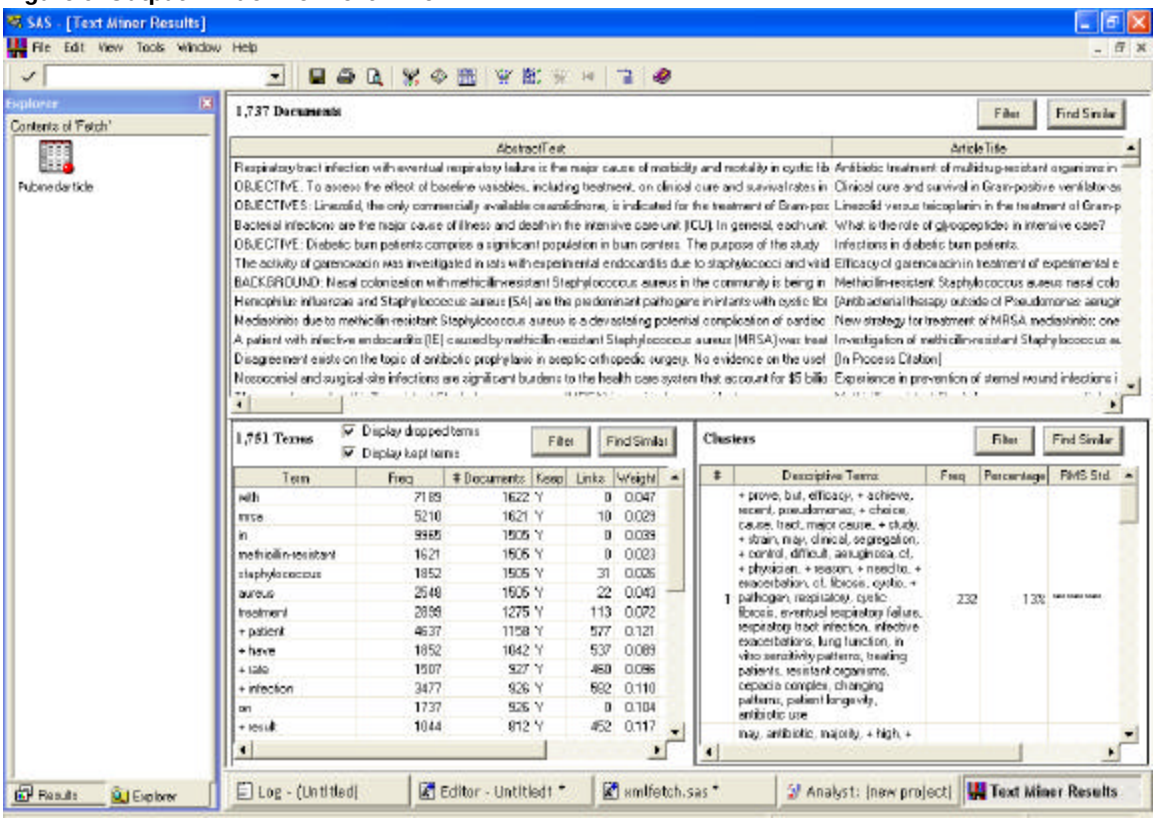


Table 1 lists the descriptors provided by the clusters in the lower right window of Figure 6.

TABLE 1. DESCRIPTORS OF TEXT MINER CLUSTERS

Cluster Number	Descriptors	Frequency
1	+ prove, but, efficacy, + achieve, recent, pseudomonas, + choice, cause, tract, major cause, + study, + strain, may, clinical, segregation, + control, difficult, aeruginosa, cf, + physician, + reason	232
2	may, antibiotic, majority, + high, + antibiotic, antibiotic prophylaxis, aseptic orthopedic surgery, noncomplex aseptic surgeries, undisputed evidence, aseptic orthopedic surgeries, osteosynthetic	116
3	early, + obtain, + concentration, cardiac, cardiac surgery, electrophoresis, gel, primary, + level, trough, due to, endocarditis, + complication, ineffective, intravenous, new, + surgery	347
4	+ rate, on, s., antibiotic, + time, + patient, colonization, + method, + objective, + associate, gram-positive, gram-positive, methicillin-sensitive, with, nasal colonization, general, diagnostic	810
5	+ compare, + mechanism, objective, retrospective, + characteristic, + prolong, increase, must, + center, nosocomial, may, major, + burn, diabetic burn patients, significant population, burn centers	116
6	resistant, but, + drug, + indicate, contrast, potential, + control, methicillin-susceptible, , + cause, not, + day, only, activity, streptococcus, + isolate, + streptococcus, + comparator	116

Of the 6 clusters, cluster 1 appears to relate more to other infections; cluster 2 relates to prevention rather than treatment. Clusters 4 and 6 appear to be the most promising to examine treatment. It is possible in Text Miner to filter the data down to cluster 4, and to re-cluster to "drill down" into the data (Table 2).

Table 2. Re-clustering of Cluster 4

Cluster Number	Descriptors	Frequency
1	+ patient, + study, + good, initial, + subset, linezolid, end, linezolid, suspected, double-blind, + randomize, design, baseline, + assess, + effect, clinical, skin, + perform, + site, including	347
2	common, major, more, as, + disease, other, many, diverse, responsible, community-acquired, hospital-acquired, + find, basis, resistance, + lineage, staphylococcal, penicillin-resistant, most, majority	115
3	more, + unit, + combination, data, most, all, + favor, intensive, + agent, death, risk, more, + disease, major, + high, different, recent, resistant, bacterial, antibiotic treatment, spread	116
4	community-acquired, no, hospital-acquired, + isolate, both, aureus, with, + compare, + method, colonization, s., + high, increasingly, + pattern, + strain, chronically, + define, but, age, general	116
5	+ rate, + infection, + cost, + burden, morbidity, + treat, + increase, may, testing, often, likely, nosocomial, increased, wound, hospitalization, not, + case, cardiac surgery, cardiac, + surgery	116

Cluster 5 appears to focus on the cost, rate, and method of infection while cluster 3 appears to focus more on treatment. The number of documents can thus be reduced from 1750 to 116.

Another means to reduce the needed examination of the documents is through the use of concept links. These can be found by right clicking on a term and scrolling down to "view concept links". Not all terms will have potential links. Note that the first time the user asks for concept links, a column appears in the term window containing the number of potential links. This column can be sorted by clicking on the name "links" so that the user can find the terms with the maximum number of links.

The concept link appears in a browser window as an html document. The link can be animated by moving the mouse cursor. All of the lines connected to words are themselves connected to related words that can be discovered by moving the mouse around. Clicking the mouse on one of the words will connect to another browser window providing the documents that comprise the link. The algorithm used to define the concept links is that of association rules. Figure 7 gives the concept links for "MRSA".

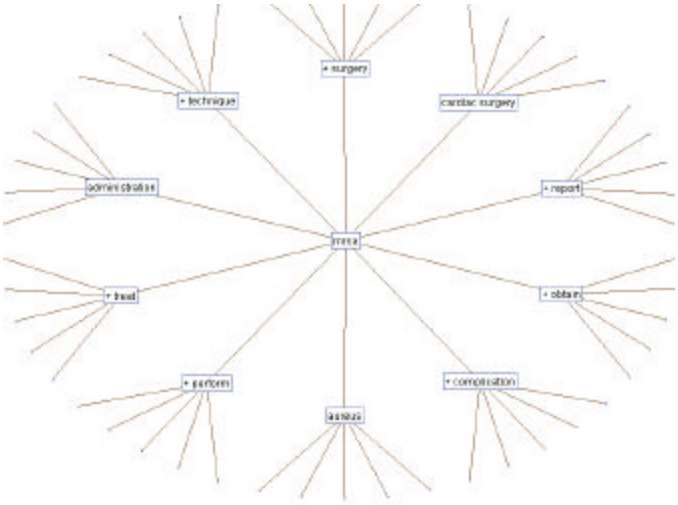


Figure 7. Concept links for "MRSA". Note that only a few of the links lead to treatment. However, of interest is the link to cardiac surgery. The links for the term "treatment" are more numerous (Figure 8).

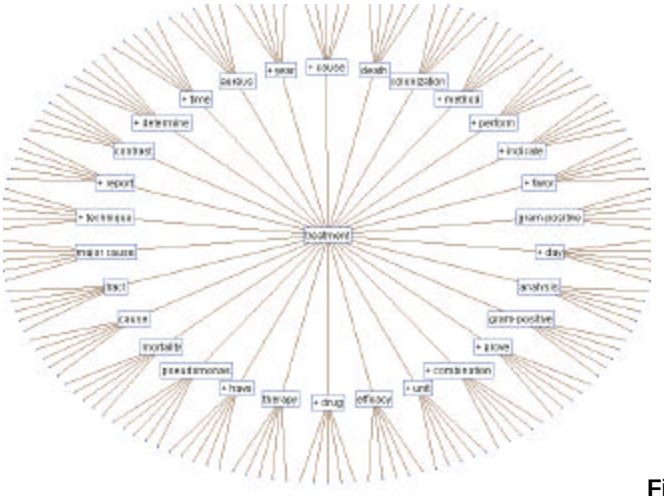


Figure 8. Concept links for "treatment". This diagram can be rotated to find the next generation of terms (Figure 9).

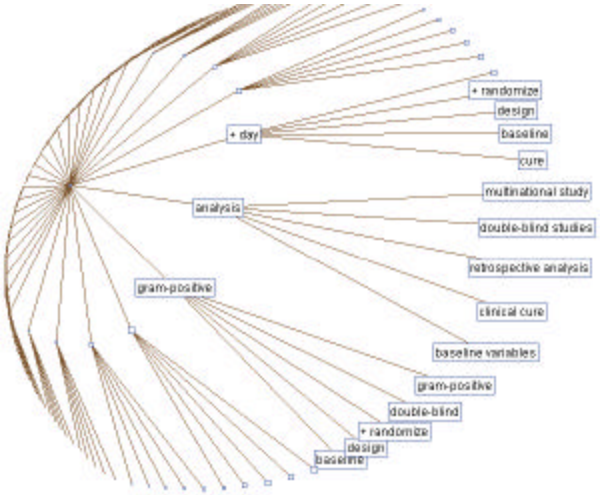


Figure 9. Rotation to the term "gram positive". This combination goes to the terms of "randomize" and "double-blind", also "retrospective analysis". It is clear that the term "gram-positive" is the most direct link to treatments for MRSA. By using the mouse to click on the diagram at the term "gram-positive" the documents that are the most directly related are quickly identified, and can be examined: OBJECTIVE. To assess the effect of baseline variab *Keywords: analysis, baseline, baseline variables, clinical cure, cure, days, design, double-blind, double-blind studies, favored, gram-positive, gram-positive, multinational study, randomized, retrospective ana*

OBJECTIVES: Linezolid, the only commercially avail *Keywords: + combination, + compare, + unit, achieved, available oxazolidinone, baseline, colonization, days, design, double-blind, efficacy, glycopeptide antibiotic, gram-positive, gram-positive, has, indicated*

Bacterial infections are the major cause of illnes *Keywords: + choice, + combination, + favor, + objective, + time, + unit, antibiotic treatment, aureus, bacteria, bacterial infections, cause, caused, compared, death, determines, fact, factors, glycopeptides, g*

The user can click on any one of the documents. However, from just a few concept links, it appears that the most recently available treatment for MRSA is the drug, "Linezolid". To validate that assumption, the links for linezolid were also examined (Figure 10).

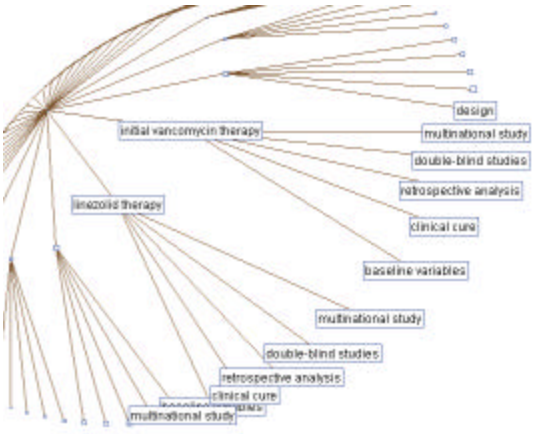


Figure 10. Concept links for "Linezolid". From these links, it appears that vancomycin is an older, more standard treatment that may not be fully successful in treating MRSA. Linezolid is a more recently developed treatment. Therefore, the investigator can concentrate on papers that deal primarily with the treatment, Linezolid and those that compare vancomycin to linezolid. The use of Text Miner can greatly reduce the time required to investigate the medical literature by allowing a quick discard of irrelevant documents, even though all were returned on a keyword search.

STANDARD MEASURES OF HEALTHCARE QUALITY

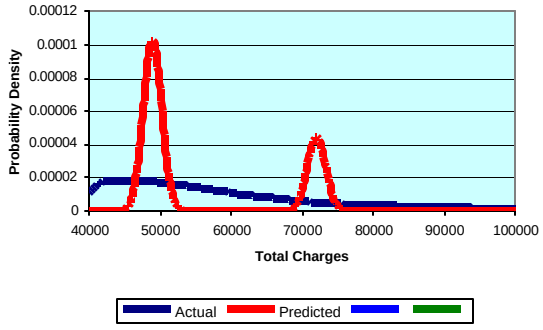
The standard procedure for examining hospital quality and cost effectiveness has been to use an equation of the form

$$y = B_0 + B_1x_1 + ... + B_kx_k + e$$

where y = dependent variable (variable to be modeled – sometimes called the response variable), $x_1, x_2, ..., x_k$ = independent variable (variable used as a predictor of y), e = random error, and B_i determines the contribution of the independent variable x_i.[14, 15] The variable y can be continuous (such as length of stay or costs), or discrete (such as mortality). Each x variable denotes the presence or absence of a risk factor for a particular patient. Therefore, y is equal to the sum of the weights B_i for each factor x_i that is positive for that patient. If y is a continuous variable such as length of stay, then the predicted LOS is the linear combination of weights. If y is a discrete variable such as mortality, an optimum threshold value is found, and mortality is predicted if the sum of the weights exceeds that threshold.

To determine a ranking, the value Expected [y (difference)]=Expected [y (predicted)-y(actual)] is computed. If Expected [y(difference)] is positive then the outcome is good; the higher the expected difference, the higher the ranking. The higher a negative expected difference, the lower the ranking. Predictions, however, tend to be linear (Figure 11). A linear model cannot predict outliers and it is the outliers that are the most costly. Patient costs tend to follow more of an exponential distribution.

Figure 11. Kernel Density Estimate for Predicted Compared to Actual Charges Using Regression



Because there are so many possible risk factors that can be included (the entire set of ICD9 codes), many are initially eliminated because they lack statistical significance in a pairwise comparison. Given a sufficiently large set of patients, almost any factor can become significant. Then a stepwise procedure is used so that the most important risk contributors are used in the model. Because there are so many possible risk factors to choose from, this procedure can result in many different models, and each commercial organization that has developed a model can use one that is different. If y is discrete, then a measure of the accuracy of the model is in the c statistics; if y is continuous, the measure is r^2. The two measures are generally relatively low in models used for predictions.

In order to create the model, an assumption has to be made that patient risk factors are uniformly entered across all providers. If one hospital regularly under-reports on risk factors, then y(predicted) will have fewer weights compared to other hospitals, and will be lower compared to other hospitals. A low predicted y will result in a low ranking. Therefore, the process rewards hospitals that tend to over-report on risk factors.

USE OF PHARMACY ORDER DATABASE AND TEXT MINER

Medication orders for individual patients can be combined in a text string. Pharmacy data entered at the point of care will be more reliable than extraction using ICD9 codes. A typical string of orders is "INSULIN LISPRO INSULIN LISPRO NITROGLYCERIN/D5W NITROGLYCERIN/D5W ENOXAPARIN SODIUM ENOXAPARIN SODIUM ASPIRIN ASP" This particular patient has diabetes and angina with mild pain. The pharmacy order database can be used as a check on manual extraction of patient risk factors as indicated by co-morbid diagnoses. Conversely, the database can be used to examine the treatment for the co-morbidities to ensure that the patients are receiving optimal care.

A diagnosis of diabetes can be made from an order of insulin. However, other diagnoses usually have multiple medications while many different diagnoses can have common medications. Congestive heart failure, in particular, includes a multitude of medications including statins, ace inhibitors, and diuretics. Albuterol can be for asthma or COPD (chronic obstructive pulmonary disease). The on-line version of Drug Facts and Comparisons contains the following information in searchable form:[16]

- The latest information on more than 850 generic and 3,700 trade name drugs
- Incremental updates for one year just by synching your handheld
- Detailed information in the following fields:

o Generic & Trade Names	o Adverse Reactions	o Contraindications	o Administration and Storage
o Pronunciation	o Precautions	o Dosing	o Assessment and

- Therapeutic class
- Action
- Dosage Forms / Strengths
- Interventions
- Overdosage
- Indications
- Lab Test Interferences
- Interactions
- Patient and Family Education

Facts and Comparisons has a searchable database. Therefore, co-morbidities related to a primary diagnosis can be used to search and the medications related to those co-morbidites will be found. As most medications have multiple uses, the presence of one medication will not necessarily be conclusive concerning a diagnosis. There are exceptions; for example, insulin does indicate Type I diabetes. A set function will be defined that relates the diagnosis X to the set containing all possible medications for diagnosis Y(X)). Then given a set Z of medications, the inverse function Z^{-1} is equal to $Y_Z=\{X_1,\ldots,X_n\}$ of potential diagnoses such that for each X_i, Z is a subset of $Y(X_i)$. If Y_Z contains only one element then it is almost certain (95%) that the patient can be assigned that diagnosis. If Y_Z contains more than only element, then the diagnoses will be flagged for manual examination. If Y_Z is empty, then there is insufficient information to define a diagnosis.

For a patient in the dataset, one text string will contain the total list of patient medications. An overlapping partition P will be defined from the text string so that for each diagnosis in the database X_i, all medications that are contained within the patient text string that are also contained in $Y(X_i)$ will define the set P_i for the patient P. For each P_i, the set Y_{Pi} will be defined to investigate the pattern of diagnosis.

Once the list of $Y(X_i)$ has been defined for a diagnosis X_i, it can be used to define a startlist in the text mining software. A startlist is a list of words pre-defined that are the only words be used to define clusters of patients with the text mining software. At the text mining initiating screen, the startlist can be specified. Then only the terms in the startlist will be retained in the text parsing. Patients with no medications from the startlist will be including in the frequency count. A one-on-one correspondence between diagnosis and startlist can be embedded into the text mining software. This will standardize the process of data searches through the patient fields. Startlists can be concatenated to look for two or more diagnoses simultaneously.

As an example of the necessary code, all medications related to diabetes were used to extract the diagnosis using the pharmacy database. Using a database of 2500 patients divided into 6 different groups, manual versus automatic extraction were compared (Table 3). The initial analysis shows that automatic extraction can identify more patient risk factors compared to manual extraction.

Table 3. Comparison of manual versus automatic extraction of diagnosis of diabetes in six groups of patients

Patient Group	Manual Extraction	Automatic Extraction
1	20.79	33.83
2	15.45	34.93
3	37.7	48.28
4	24.73	31.36
5	12.4	35.14
6	15.28	31.58

In each group, automatic extraction using the medications for diabetes results in a statistically significant difference in documenting diabetes. Automatic extraction requires little time and will allow quick flaging to ensure proper documentation and billing.

While a text string containing all generic names of medications per patient may seem very simple to obtain, it is worth pointing out that the transformation from the original format to the desired format turns out to be outside of the scope of relational databases, as well as most automatic data treatment tools. In a relational database, the original records (called rows or tuples) have the information in fields (called columns or attributes). Combining information from several records is possible in Structured Query Language (SQL). However, it is not possible to carry out the desired transformation in SQL, as it basically involves combining values in a single column (the same field in several records) into a single value (the concatenation of all strings into a single string). This transformation is carried out in SQL through the use of aggregate functions. Unfortunately, all the aggregate functions available in SQL (min, max, sum, count and average) are numerical, and cannot process strings in the desired way. For the purposes of the experiment, therefore, a program was written that performed the following two steps: first, all records in the Pharmacy database were clustered by patient identifier; second, within each cluster, all the generic name field values were concatenated into a single string with a comma separator.

For example, patients can be identified by the number of medications prescribed that reduce cholesterol levels, and the number of physicians involved in the prescribing. In order to do this, the medications have to be identified by initial order date as well as the date at which the medication was discontinued to determine whether the medication was changed, or whether a second medication was prescribed that has similar properties of an initial medication. As Table 4 indicates, the analysis divided the patients into 6 clusters.[17]

Table 4. Pain Medications by Clustering of Patients

Cluster	Antibiotics
1	Cefazolin, vancomycin, mupirocin
2	Cefazolin
3	Ciprofloxacin
4	Promethazine, hydrocodone, meperidine, trazodone
5	Vancomycin, mupirocin
6	None

The consultant pharmacist was able to rank them into different levels of pain intensity using the medications prescribed.

Table 5. Antibiotics by Clustering of Patients

Cluster	Severity Ranking	Pain Medications
1	6	Acetaminophen, morphine, temazepam, famotid, temazepam, haloperidol, lorazep
2	1	Fentanyl, morphine sulfate, midazolam, famotidine
3	5	Propoxyphene
4	2	Promethazine, hydrocodone, meperidine, trazodone
5	4	Zolpidem, oxycodone, docusate, alprazolam, lansoprazole, acetaminophen, morphine sulfate, temazepam, famotid
6	3	temazep

Similarly, the antibiotics used by each natural grouping of patients is given in Table 5. Note that cluster 4 has more intense use of antibiotics while ranking second in use of pain medications. Figure 12 gives the overall usage of antibiotics. By cross-referencing the prescription use with patient outcomes, the appropriateness of prescribing can be examined (Figure 13).

Figure 12. Number of Patients Prescribed Antibiotics

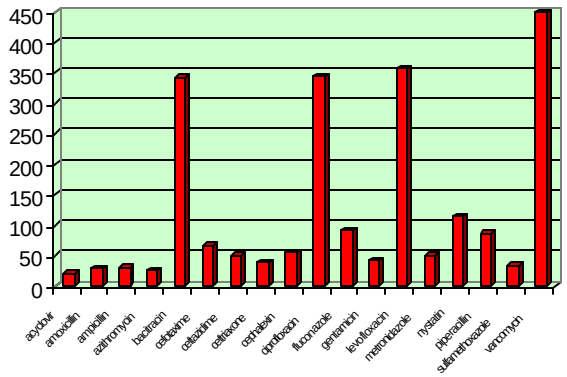
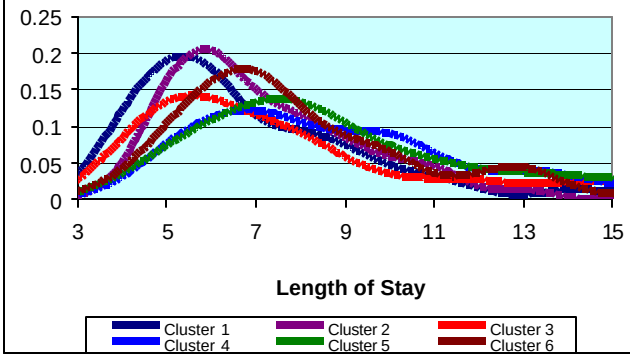
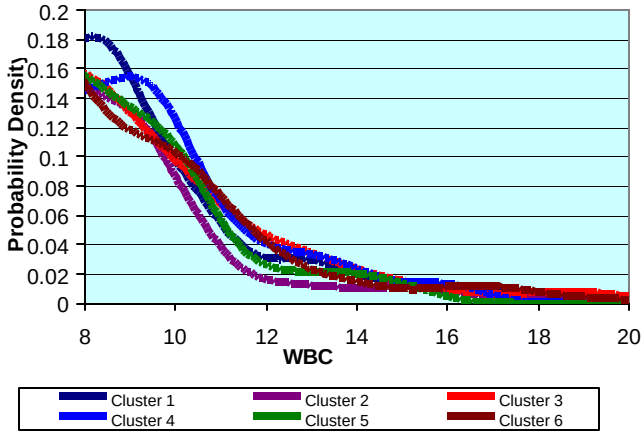


Figure 13. Length of Stay by Patient Cluster



Note that patients in clusters 1 and 2 are more likely to have shorter stays compared to patients in clusters 4 and 5. There is a natural ranking of severity in the six clusters. Patients in cluster 1 (with the least need for pain medication) have the highest probability of a stay of less than 5 days; patients in cluster 4 are the least likely to be discharged within 5 days and the highest probability of staying 10 or more days.

Figure 14. White Blood Count by Cluster



Ejection fraction measures the elasticity of the heart with 45-50% as normal, and values < 40% defining congestive heart failure. Many of the patients have ejection fractions less than 20%. Cluster 5 has the lowest proportion of patients with a normal ejection fraction while cluster 3 has the highest. The ranking of risk is in the order 3<4<1<2,5<6. For elevated white blood cell count (Figure 14) has ranking 1,4<2,3,5,6.

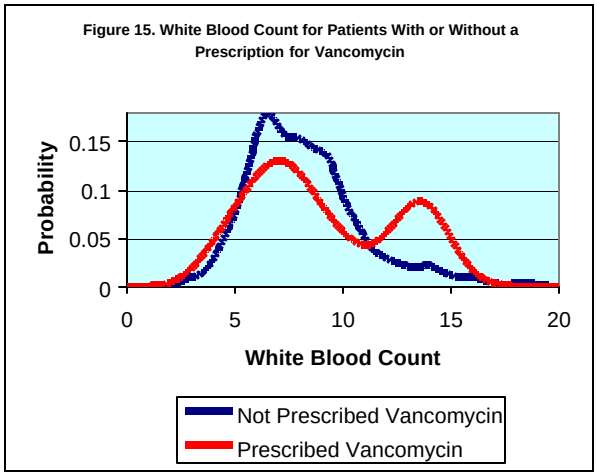
For hospital length of stay, clusters 1,2,3 have the shortest stay with clusters 4,5,6 the longest. The ordering is 1<2<3<6<5<4. Although there is not a statistically significant difference in survival (inpatient), there is a ranking of this as well: 2<1<3<4<6<5. It becomes clear from these rankings that patients in clusters 1 and 2 generally have the highest risk with patients in clusters 3 and 4 at a lower level of risk; patients in clusters 5 and 6 have the lowest risk level. Therefore, it is possible to define a risk adjustment based upon medications.

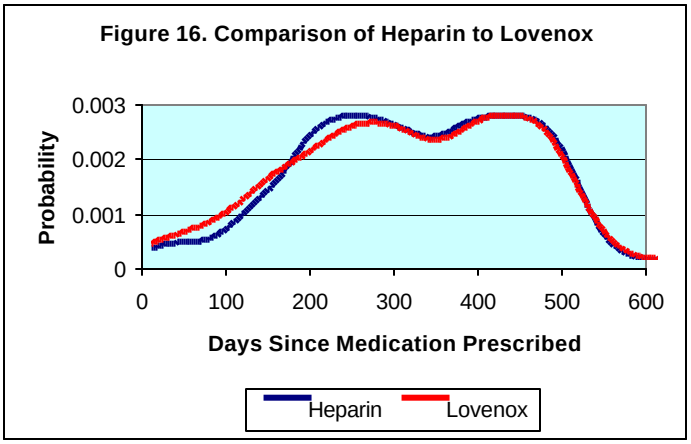
PRESCRIBING PRACTICES

Patients treated with Vancomycin were examined. The American Academy of Orthopaedic Surgeons has issued an advisory statement with the following recommendation: Vancomycin should be reserved for the treatment of serious infection with beta-lactam-resistant organisms or for treatment of infection in patients with life-threatening allergy to beta-lactam antimicrobials.[18-20] Similarly, the Hospital Infection Control Advisory Committee recommends:[21]

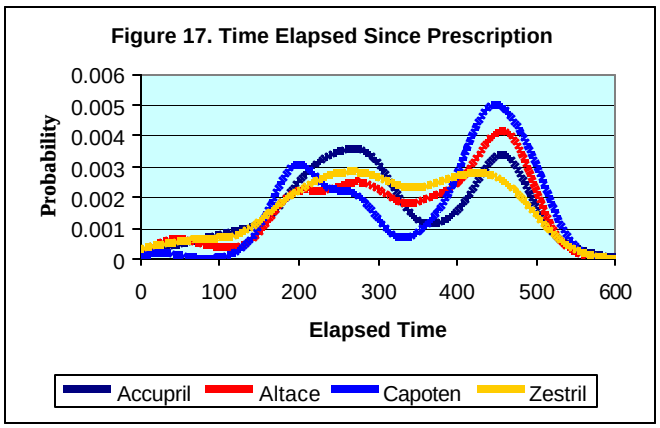
> "Cefazolin provides adequate coverage for many clean-contaminated operations. If a patient cannot safely receive a cephalosporin because of allergy, a reasonable alternative for gram negative coverage is aztreonam. However, an agent such as clindamycin or metronidazole should also be included to ensure anaerobic coverage."

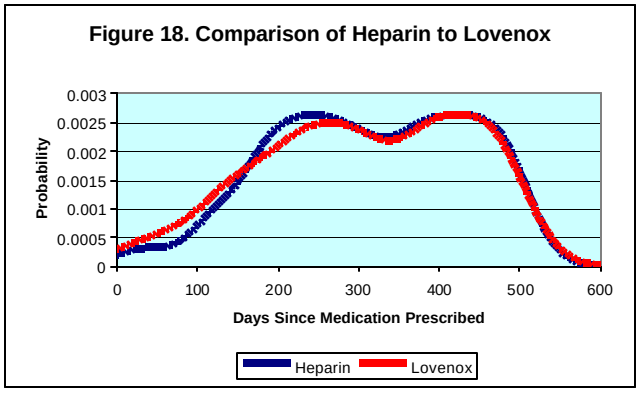
Use of Vancomycin prophylactically in combination with cefazolin is not more effective over the use of cefazolin alone.[22-24] There remains substantial inappropriate use of vancomycin.[25, 26] More appropriate use does yield considerable cost savings.[27] The formulary can be changed to reduce the inappropriate use of Vancomycin.[12, 28] All patients with Vancomycin prescriptions should have some indication of infection and should have a culture and sensitivity test. They should have elevated white blood cell counts.[29] Patients on Vancomycin were compared to those without in terms of their white blood cell count (Figure 15). Note that for patients with Vancomycin, the white blood count mostly followed the same values compared to patients without Vancomycin.





There is only a small group of patients on Vancomycin with white blood cell counts greater than 10, demonstrating that most of the Vancomycin prescriptions were for prophylactic use of the antibiotic. In this case, alternative prophylactic antibiotics should be proposed to physicians. This is in contrast to the white blood cell count in relationship to Cipro (Figure 16). The Cipro prescribed was primarily used prophylactically with no peak beyond the one at 8 units. Consider, for example, ace inhibitors used for heart patients, and for patients with diabetes (Figure 17). There is a decided decline in the use of the drugs, Altace and Capoten in favor of the drug, Accupril. The shift started approximately 200 to 300 days from time zero.

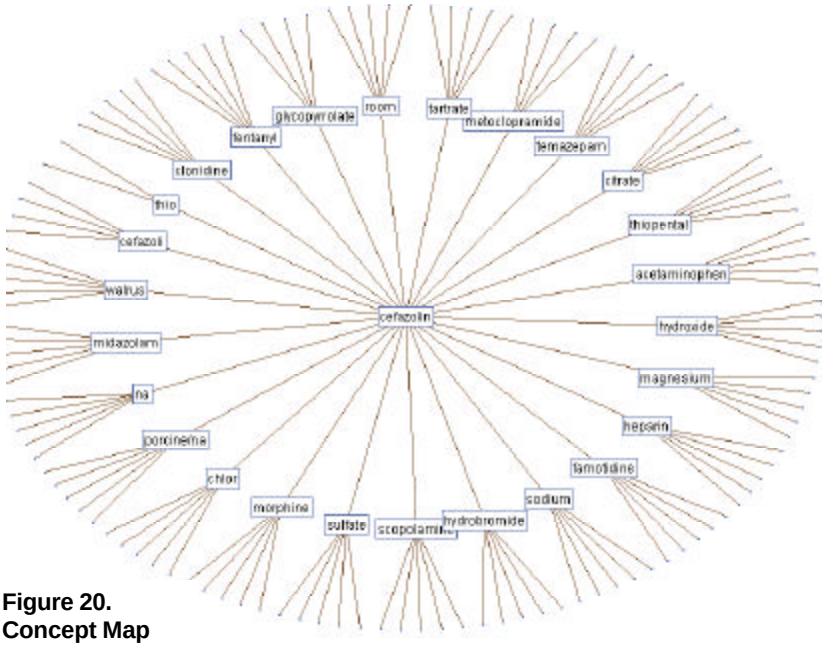




A similar analysis was performed to examine the likelihood of a shift from Heparin to Lovenox in the same population (Figure 18). Until the value of approximately 150 days since surgery the curves are relatively the same. However, there is a recent shift away from Heparin toward the use of Lovenox. Followup will determine whether the shift continues.

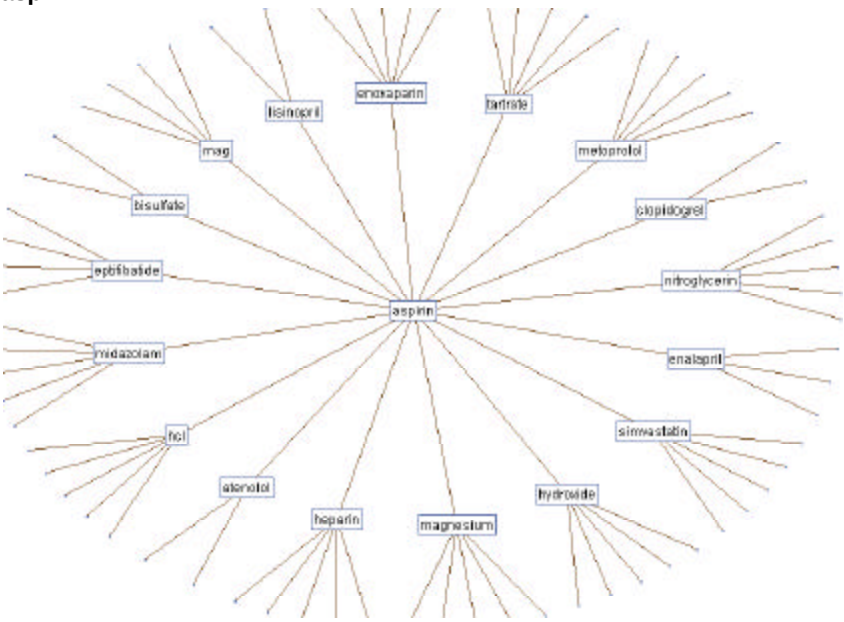
With the pharmacy database, it is possible to examine combinations of prescribing habits using concept links. The antibiotic, cefazolin is standard protocol for patients in open heart surgery. Therefore, the medication links with many different medications also used for patients in open heart (Figure 19). The medication, aspirin, is not used for all patients, and some meaningful relationships can be identified (Figure 20). Aspirin is frequently given to patients with a history of heart attack, as are the drugs, clopidogrel and nitroglycerin. Nitroglycerin has even fewer links to other medications, indicating that fewer patients have a prescription for it (Figure 21). Insulin also has only a few associations (Figure 22).

Figure 19. Concept Map for "cefazolin"



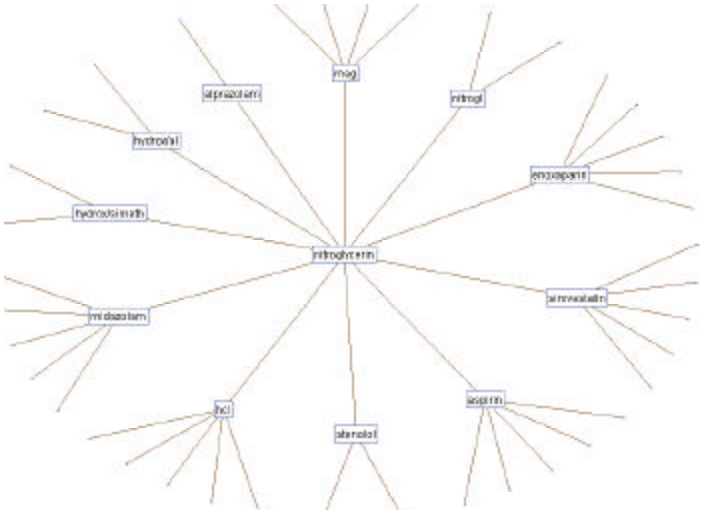
Cefazolin is often used prior to surgery to prevent infections. Since its use is almost automatic, the number of other medications that link are considerable. In particular, it is linked to morphine, almost always the initially prescribed pain medication after open heart surgery. Similarly, famotidine is routinely prescribed to prevent stomach problems resulting from the ingestion of drugs.

Figure 20. Concept Map for "aspirin"



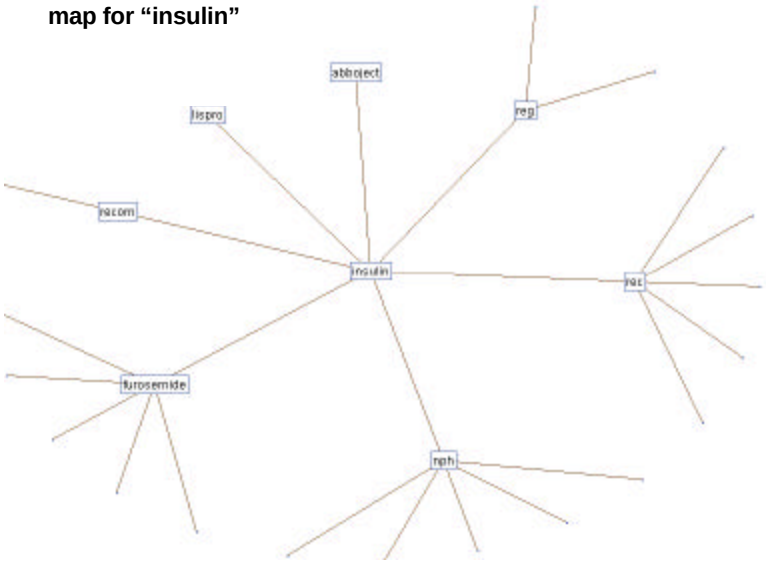
Aspirin is given to patients who have suffered a heart attack. Its purpose is to thin the blood and prevent further clots. It relates to both clopidogrel and nitroglycerin, also given to patients with heart attacks.

Figure 21. Concept map for "nitroglycerin"



As Figure 21 also demonstrates, nitroglycerin and aspirin are symmetrically related. It is also related to midazolam and alprazolam, which have similar properties. . It is possible to perform a crawl through PubMed (www.pubmed.gov) to examine the relationship between the medications. A crawl returned a total of 27 documents using the keyword search "nitroglycerin+midazolam". The returned document can be examined using Text Miner as well. Of the 27 documents, 10 clustered into a group concerning heart problems, particularly angioplasty. Another 15 examined the means of delivery, including injection. The remaining 2 papers examined the impact of nitroglycerin on the heart mechanism.

Figure 22. Concept map for "insulin"



Most of the connections to insulin are for specific types of insulin (nph, reg, lispro) or for the method of deliver (abboject). The remaining medication (furosemide) is a diuretic with several purposes. This connection, too, can be investigated using Text Miner. A search of "insulin+furosemide" returned 100 documents. The initial clustering was scattered, and the number of clusters was restricted to a maximum of 5. Of the 100, 21 papers cluster on the need of treatment of kidney problems using furosemide for patients with diabetes. Another 21 papers discuss patients who develop diabetes. A total of 51 focus on clinical trials testing the medication furosemide, and the remaining 7 discuss the medical properties.

USE OF ADMINISTRATIVE DATA AND TEXT MINER

For each patient, all secondary codes are combined into one text string. An example of such a string is "4271 42731 42781 4019 41401 412 2724". This patient then suffers from unspecified paroxysmal tachycardia, atrial fibrillation, cardiac dysrhythmia, unspecified essential hypertension, coronary atherosclerosis, old myocardial infarction, and lipoid metabolism disorder. A second patient has "4271 412 4280 2724 4019 27800" or unspecified paroxysmal tachycardia, old myocardial infarction, congestive heart failure, lipoid metabolism disorder, unspecified essential hypertension, and obesity. These two examples demonstrate that certain ICD9 codes tend to be associated. Because of the manner in which the codes are defined, it becomes possible to examine these defined text strings for patterns.[30]

Data were used from two different sources and restricted to heart DRGs (Table 6). The data were collected from a total of 13 hospitals within Kentucky. To validate the results, the rankings resulting from the text clusters were compared to those provided by Solucient, Inc. (Evanston, Ill.) To continue the validation process, a second dataset containing the same DRGs and a second set of 8 hospitals were scored using the text clusters from the original dataset (Table 7).

Table 6. Frequency of DRG codes for dataset 1.

DRG Code	Frequency
104	1125
105	1396
106	252
107	5410
108	586
109	4405
110	1560
Total	14734

Table 7. Frequency of DRG codes for dataset 2.

DRG Code	Frequency
104	747
105	1028
106	2034
107	2137
108	222
109	0
110	913
Total	7081

Table 8. Clusters identified using Text Mining, along with most frequent codes found in the clusters

Cluster Number	Frequency	Label
1	1682	Mild general risk factors
2	3187	More severe general risk
3	1139	Complications after surgery
4	1133	Unrelated risk factors and aortic problems
5	1469	IDDM Diabetes with complications
6	907	Moderate risk with specific factors
7	4159	Severe risk and severe complication after surgery
8	586	Very severe complications after surgery
9	1856	Severe complications after surgery

The defined text strings were clustered using the expectation maximization algorithm using the data identified in Table 4. A total of 9 clusters were identified (Table 9).

There is a natural ordering in terms of severity of risk factors: 2<4<6<5<1<7<3<8<9 by considering average total charges and 2<6<4<5<1<3<7<9<8 by considering length of stay. However, there are only small differences between some of the clusters that can lead to natural groupings as well.

The clusters 2,4,6 can be grouped together as can 8,9. Then 3 and 7 can be grouped leaving 1,5 to create four categories of risk, A(2,4,6)<B(1,5)<C(3,7)<D(8,9). Combining text clusters gives 4 different categories of patient severity, a fairly standard number. Then the overall relationship of rank to total charges and length of stay are given in Figures 23,24.

Figure 23. Kernel Density Estimate for Length of Stay by Rank

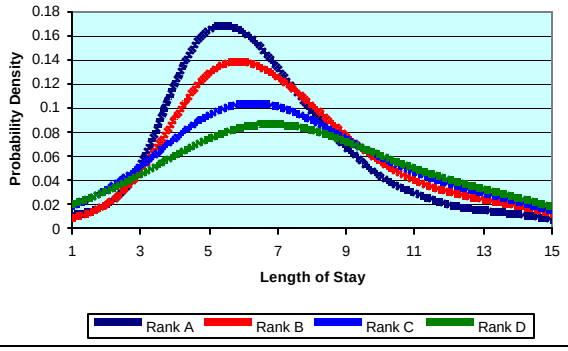
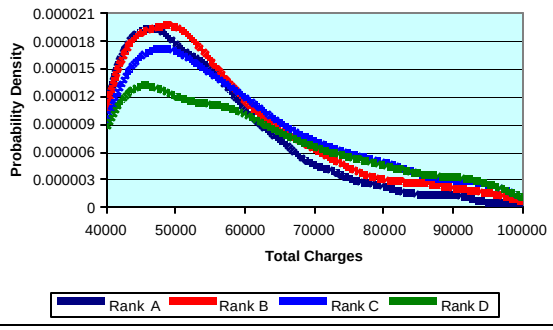
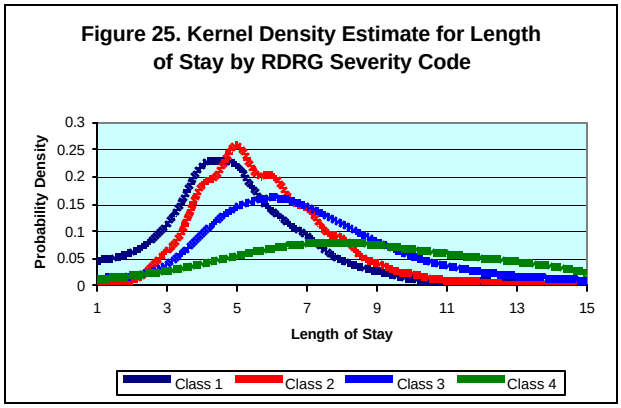
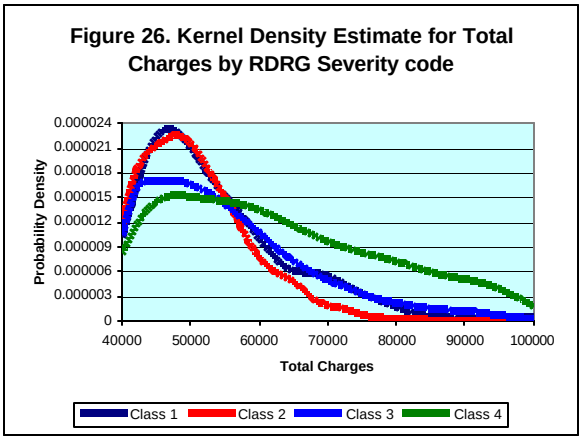


Figure 24. Kernel Density Estimate for Total Charges by Rank



Note that as the rank increases, the probability that the length of stay goes beyond 9 days is greater. Similarly, the probability that total charges exceeds 60,000 increases as the rank increases. It would seem reasonable that patients at higher risk will stay longer at increased cost. This is not true for standard Solucient RDRG category (Figures 25,26).





In this case, it is more likely that class 2 patients will have lower total charges compared to class 1 patients; moreover, class 1 patients have almost the same total charges as class 3 patients beyond the 55,000 threshold.

Table 9 compares the 13 hospitals to determine the proportion of patients in each cluster. The difference is statistically significant ($p<0.0001$). It clearly demonstrates that the hospitals are reporting at different levels of digits of the ICD9 codes. Hospital 1 reports more risk factors compared to hospitals 4 and 9, and will be penalized in quality rankings.

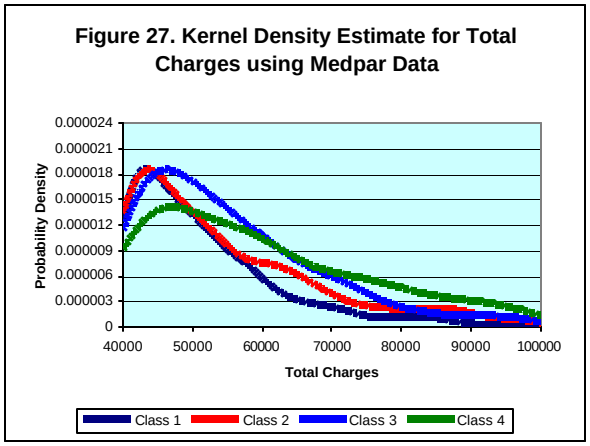
Table 9. Proportion of patients in each cluster by hospital in dataset 1

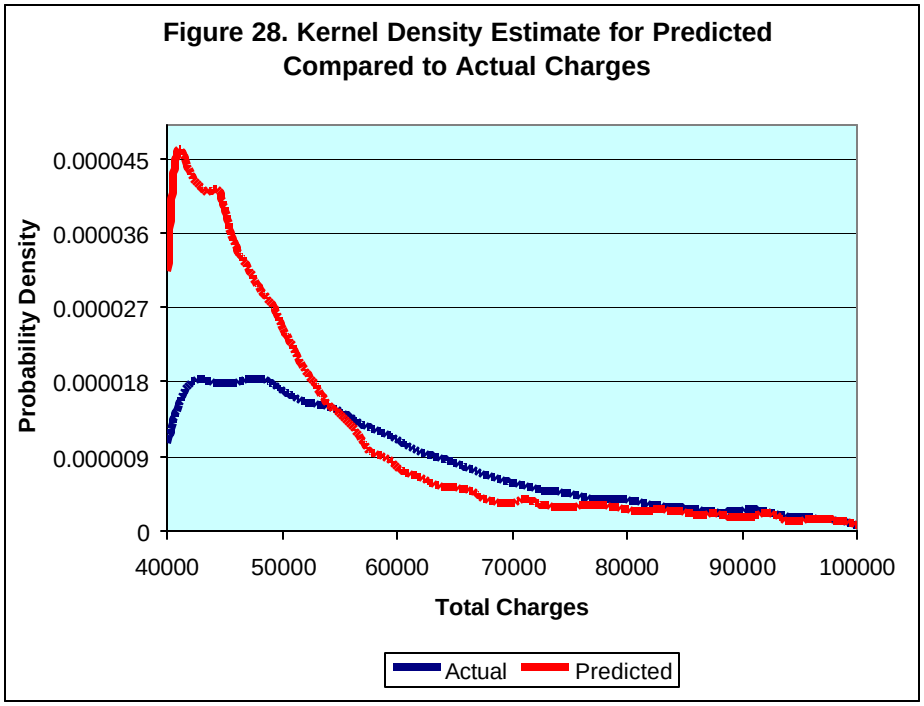
Hospital Number	1	2	3	4	5	6	7	8	9
1	9.8	22.3	6.5	4.4	7.9	4.5	20.6	3.6	20.5
2	6.7	25.4	5.4	6.7	10.8	4.5	23.2	2.8	14.5
3	15.1	12.9	8.1	5.2	7.7	10.0	17.0	4.4	19.6
4	14.5	18.8	9.2	6.8	7.3	7.2	25.2	2.5	8.5
5	10.9	17.3	11.5	9.3	9.1	6.3	21.6	3.9	10.0
6	14.9	16.5	5.5	6.2	10.5	5.3	20.6	2.6	17.7
7	11.9	22.2	8.6	5.5	11.1	6.5	24.9	1.3	8.0
8	9.6	18.4	12.4	4.9	7.8	7.8	28.5	0.8	9.8
9	8.1	18.4	4.2	9.5	11.5	4.2	28.8	8.3	7.1
10	11.6	18.6	6.7	7.8	6.6	10.5	27.4	1.6	9.1
11	11.5	26.6	6.5	9.4	13.2	7.3	16.1	1.0	8.4
12	8.2	15.6	6.7	8.0	6.3	2.6	33.3	2.2	17.1
13	13.1	23.3	6.4	6.8	7.1	5.3	25.9	2.4	9.7

VALIDATION

SAS Text Miner has a scoring mechanism that can be used to cluster new data using the clusters developed in the original dataset. Therefore, it is possible to validate the initial results using a second dataset.

Using the second set of data, containing 7081 patients in 8 hospitals to demonstrate the effectiveness of the text mining process, the four ranks defined through text mining yielded a relationship showing that total charges increase as the rank increases (Figure 27)





The scoring mechanism of data mining was used to determine whether the model could predict the total charges compared to the actual reported total charges (Figure 28). The scoring code is given in Figure 29. Once the scoring is completed, there are two datasets defined: emdata.td_xxxx and emdata.sd_xxxx. These datasets contain the scoring information that was used to construct Figure 28.

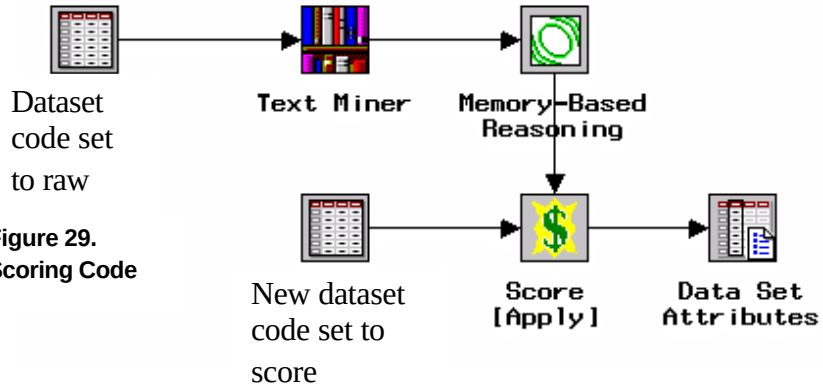


Figure 29. Scoring Code

CONCLUSION

It is possible to develop a model to rank the quality of care so that the model does not assume uniformity of data entry. The model can also be validated by examination of additional data. The means of developing the model is to use the stemming properties of the ICD9 codes where the first three digits of the code represent the primary category while the remaining two digits represent a refinement of the diagnosis. The model compares well to those developed through the standard logistic regression technique.

Hospitals that rank low should compare their coding using text analysis to determine where the coding can be improved by shifting ranks from low to high. Text analysis provides means for hospitals to examine their own coding practices.

Many of the techniques developed for healthcare can be used in other businesses to examine relationships in complex data.

REFERENCES

1. Herdman DBaR. How Does the VA National Formulary Compare with Private Insurance Formularies for Drugs and Devices and with Other Government Formularies? *VA Pharmacy Formulary Analysis Committee*. Available at: www.nap.edu/html/VA_national_formulary/ch5.pdf, 2003.

2. Lautenbach E, Larosa L, Kasbekar N, Peng H, Maniglia RJ, Fishman NO. Fluoroquinolone utilization in the emergency departments of academic medical centers: prevalence of, and risk factors for, inappropriate use. *Archives of Internal Medicine.* 2003;163(5):601-605.

3. Gross R, Morgan A, Kinky D, Weiner M, Gibson G, Fishman N. Impact of a hospital-based antimicrobial management program on clinical and economic outcomes. *Clinical Infectious Diseases.* 2001;33(3):289-295.

4. Minooee A, Rickman L. Expanding the role of the infection control professional in the cost-effective use of antibiotics. *American Journal of Infection Control.* 2000;28(1):57-65.

5. Hecker M, Aron DC, Patel NP, Lehmann MK, Donskey CJ. Unnecessary use of antimicrobials in hospitalized patients: current patterns of misuse with an emphasis on the antianaerobic spectrum of activity. *Archives of Internal Medicine.* 2003;163(8):972-978.

6. Cunha B. Quinolones: clinical use and formulary considerations. *Advances in Therapy.* 1998;15(5):277-287.

7. Empey K, Rapp R, Evans M. The effect of an antimicrobial formulary change on hospital resistance patterns. *Pharmacotherapy.* 2002;22(1):81-87.

8. Franks P, Fiscella K. Effect of patient socioeconomic status on physician profiles for prevention, disease management, and diagnostic testing costs. *Medical Care.* 2002;40(8):717-724.

9. Smith D. Decreased antimicrobial resistance following changes in antibiotic use. *Surgical Infections.* 2000;1(1):73-78.

10. Sommers B. Economics of antibiotic administration. *Critical Care Nursing Clinics of North America.* 2003;15(1):89-96.

11. Rapp R, Ribes J, Overman S, Darkow T, Evans M. Annals of Pharmacotherapy. *36.* A decade of antimicrobial susceptibilities at the University of Kentucky Hospital;4(596-604).

12. Gunderson B, Ross G, Ibrahim K, Rotschafer J. What do we really know about antibiotic pharmacodynamics. *Pharmacotherapy.* 2001;21(11 Pt 2):302S-318S.

13. Vlahovic-Palcevski V, Morovic M, Palcevski G. Antibiotic utilization at the university hospital after introducing an antibiotic policy. *European Journal of Clinical Pharmacology.* 2000;56(1):97-101.

14. Iazzoni L. *Risk adjustment for measuring healthcare outcomes*. 2nd ed. Chicago: Healthcare Administration Press; 1997.

15. O'Keefe K. Accounting for severity of illness in acutely hospitalized patients: a framework for clinical decision support using DYNAMO. *General Electric Medical Systems.* Available at: http://www.gemedicalsystems.com/inen/prod_sol/hcare/resources/library/article07.html, 2003.

16. Anonymous. A to Z Drug Facts with Auto-Updates. *Facts and Comparisons*. Available at: http://www.unboundmedicine.com/news_um_dfs.html, 2002.

17. Cerrito P, Badia A, Cerrito J, Cox J. Use of Text Miner to Automatically Abstract Patient Information from the Pharmacy Order Database. In: Pharmasug, ed. *Pharmasug Proceedings*. Miami: SAS Institute, Inc.; 2003:365-370.

18. Anonymous. Advisory Statement. *American Academy of Orthopaedic Surgeons*. Available at: http://www.aaos.org/wordhtml/papers/advistmt/vancomyc.htm, 2003.

19. Anonymous. ACC/AHA guidelines for coronary artery bypass graft surgery. *Circulation.* 1999;100(13):1464-1480.

20. Anonymous. ASHP therapeutic guidelines on antimicrobial prophylaxis in surgery. *American Journal of Health System Pharmacy.* 1999;56(18):1839-1888.

21. Mangram A, Horan T, Pearson M, Silver LC, Jarvis W. Guideline for Prevention of Surgical Site Infection. *American Journal of Infection Control.* 1999;27(2):97-134.

22. Niederhauser U, Vogt M, Vogt P, Genoni M, Kunzli A, Turina M. Cardiac surgery in a high-risk group of patients: is prolonged postoperative antibiotic prophylaxis effective? *Journal of Thoracic & Cardiovascular Surgery.* 1997;114(2):162-168.

23. Finkelstein R, Rabino G, Mashiah T, et al. Vancomycin versus cefazolin prophylaxis for cardiac surgery in the setting of a high prevalence of methicillin-resistant staphylococcal infections. *Journal of Thoracic & Cardiovascular Surgery.* 2002;123(2):326-332.

24. Salminen U, Viljanen T, Valtonen V, Ikonen T, Sahlman A, Harjula A. Ceftriaxone versus vancomycin prophylaxis in cardiovascular surgery. *Journal of Antimicrobial Chemotherapy.* 1999;44(2):287-290.

25. Thomas A, Cieslak P, Strausbaugh L, Fleming D. Effectiveness of pharmacy policies designed to limit inappropriate vancomycin use: a population-based assessment. *Infection Control & Hospital Epidemiology.* 2002;23(11):683-611.

26. Richardson L, Wiseman S, Malani P, Lyons M, Kauffman C. Effectiveness of a vancomycin restriction policy in changing the prescribing patterns of house staff. *Microbial Drug Resistance.* 2000;6(4):327-330.

27. Phillips E, Louis M, Knowles S, Simor A, Oh P. Cost-effectiveness analysis of six strategies for cardiovascular surgery prophylaxis in patients labeled penicillin allergic. *American Journal of Health System Pharmacy.* 2000;57(4):339-345.

28. Quale J, Landman D, Saurina G, Atwood E, DiTore V, Patel K. Manipulation of a hospital antimicrobial formulary to control an outbreak of vancomycin-resistant enterococci. *Clinical Infectious Diseases.* 1996;23(5):1020-1025.

29. Young LY, Holland EG. Interpretation of Clinical Laboratory Tests. In: Koda-Kimble MA, ed. *The Clinical use of Drugs*. 6th ed. Vancouver: Applied Therapeutics, Inc.; 1995:4=1 to 4-20.

30. Fischer W. A comparison of PCS construction principles of the American DRGs, the Austrial LDF system, and the German FP/S E system. *Casemix.* 2000;2(1):12-20.

CONTACT INFORMATION

Patricia B. Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40214
Work Phone: 502-560-8534 Fax: 502-852-7132 Email:pcerrito@louisville.edu