Paper SD07

Using SAS to Make an Independent Assessment of Electronic Medical Records

Patricia B. Cerrito, University of Louisville and, Louisville, KY

ABSTRACT

There are many claims concerning the benefits of electronic medical records: reduction in medical errors, more accurate billing records, more timely treatment of patients. However, these claims are rarely validated through statistical analysis. Using different types of data in the electronic medical database, it is possible to investigate the validity of claims of improvement in the quality of patient care. It is also possible to investigate the impact of physician decision-making on patient outcomes. For the ER, it is also possible to examine the relationship of triage (the ranking of severity of patient complaint) to patient ER discharge (home, home health, physician follow-up, admission to hospital).

Kernel density estimation (PROC KDE) and Text Miner will be used in conjunction with more inferential statistical methods to investigate data from electronic medical records based in a hospital emergency room. Initial data collected included patient histories, physician orders, laboratory results, patient throughput records, and billing codes. In particular, variability in physician decisions, patient co-morbidities, and outcomes will be investigated.

INTRODUCTION

With the availability of an electronic medical record, it is possible to examine issues of process that can both reduce cost and improve patient care. The IBEX system (Ibex HealthSystems, Inc.; Rosement, IL.) records patient length of stay (LOS) in the emergency room (ER). This is defined as time between triage and discharge. Statistical analysis can be used to examine differences across physicians and nurses in terms of LOS.

Data from a total of 6 physicians and 19 RNs were collected from a total of 1000 patients from January ñ March of 2004. Of that total, 54% were male with an average age of 39 years (±19 years). There was a statistically significant difference in age by gender ($p<0.0001$) with the average age of females equal to 42 and the average of males equal to 36. Neither age nor gender was statistically significant by physician or RN.

A total of 43 patients were admitted to the hospital while 38 were transferred and 3 were admitted to a nursing home. Out of the 1000, 904 were discharged home with 7 receiving home health services. Another 8 eloped (left before discharge). The remaining patients did not have a recorded disposition.

RESULTS OF ANALYSIS

DATA VISUALIZATION USING KERNEL DENSITY ESTIMATION

Kernel density estimation remains an extremely important data visualization technique. The kernel density estimate is defined by the equation:

$$\hat{f}(x) = \frac{1}{na_n}\sum_{j=1}^{n} K\left(\frac{x - X_j}{a_n}\right)$$

where n is the sample size, K is a known density function, and a_n is a constant depending upon the size of the sample that controls the amount of smoothing in the estimate. Note that for most standard density functions K, where x is far in magnitude from any point X_j, the value of $K\left(\frac{x - X_j}{a_n}\right)$ will be very small. Correspondingly, where many data points cluster together, the value of $\hat{f}(x)$ will be high. K can be the standard normal density or the uniform density. Simulation studies have demonstrated that the value of K has a very limited impact on the value of the density estimate. The value of the bandwidth, a_n, however, has a substantial impact on the value of the density estimate. The true value of this bandwidth must be estimated, and there are several methods available to optimize this estimate.

PROC KDE uses only the standard normal density for K but allows for several different methods to estimate the bandwidth, as discussed below. The default for the univariate smoothing is that of Sheather-Jones plug in (SJPI):

$$h = C_3\left\{\int f''(x)^2\,dx, \int f'''(x)^2\,dx\right\}C_4(K)h^{5/7}$$

where C_3 and C_4 are appropriate functionals. The unknown values depending upon the density function f(x) are estimated with bandwidths chosen by reference to a parametric family such as the Gaussian as provided in Silverman:

$$\int f''(x)^2dx = \sigma^{-5}\int \phi''(x)^2dx \approx 0.212\sigma^{-5}$$

However, the procedure uses a different estimator called the simple normal reference (SNR) as the default for the bivariate estimator:

$$h = \hat{\sigma}\left[4/(3n)\right]^{1/5}$$

along with Silvermanís rule of thumb (SROT):

$$h = 0.9\min[\hat{\sigma}, (Q_1 - Q_3)/1.34]n^{-1/5}$$

and the over-smoothed method (OS):

$$h = 3\hat{\sigma}\left[1/70\sqrt{\pi}n\right]^{1/5}$$

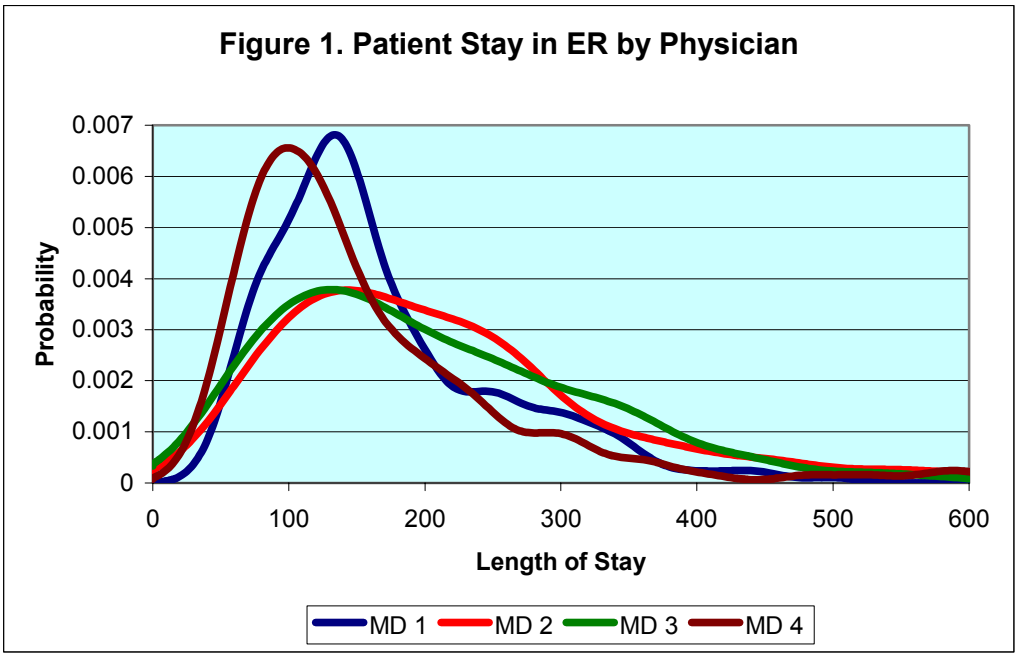
PROC KDE will accept one or two variables, depending upon whether the density is univariate or bivariate. The density estimate is stored in an output file containing the values of x from gridl (minimum) to gridu (maximum). The default number of data points is 401. This can be changed with the option ng=*number*. In addition, the default bandwidth estimate can be altered with the option bwm=*number*. The number is a multiple of the default bandwidth.

PROC KDE has been changed in version 9.1 from version 8.2 so that it can accommodate multiple distribution statements, and also provides the option of graphical display. However, the great advantage of using PROC KDE is the ability to overlay different segments of the population to compare probability distributions. Therefore, SAS/Graph should still be used to compare populations; this feature is not available in PROC KDE. The general code is

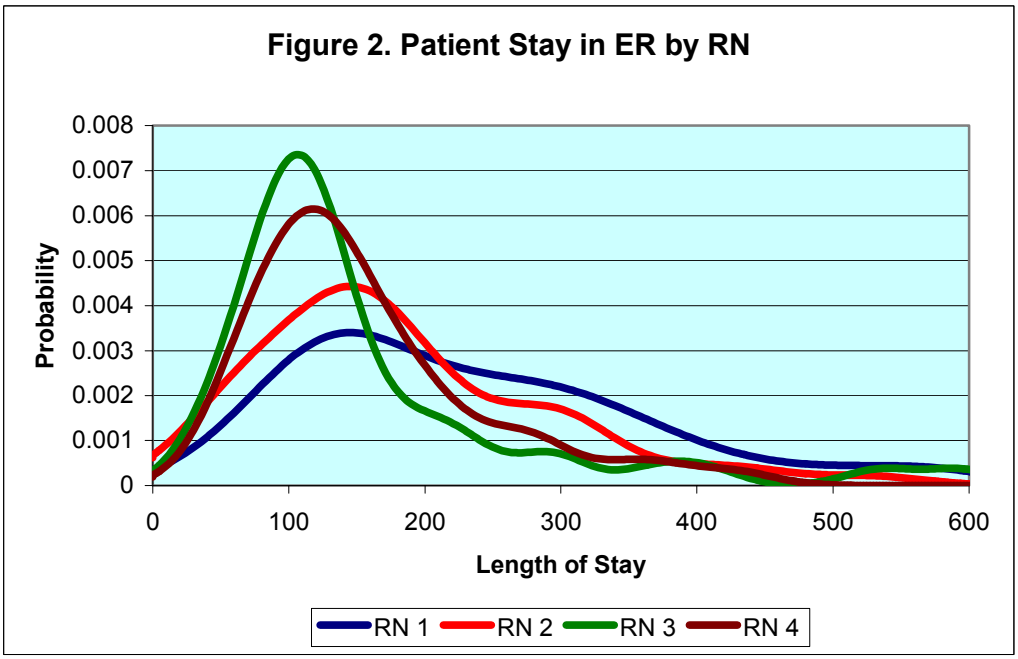
```
PROC KDE data=work.sample;
Univar LOS/gridl=0  gridu=600 out=outkde;
By MD;
Run;
```

The procedure still assumes that any variable used in the BY statement has first been sorted.

There was a statistically significant difference in length of stay (as recorded in total number of minutes from admission to discharge) by physician ($p<0.0001$) and by RN ($p<0.0001$). The highest average by physician was equal to 222.46 minutes compared to the lowest of 157.85 for a difference in range of 64.88 minutes. That hour difference increases costs while reducing the quality of care. The length of stay ranged from 23 minutes to 935 minutes. To demonstrate the difference in time, kernel density estimation was used to examine the entire time distribution by physician (Figure 1).



MDs 1 and 4 have a high peak around 100,120 minutes while MDs 2 and 3 show considerably more variability with a high likelihood that patients will stay 250 or more minutes in the ER. A similar range of times and differences occurs across RNs ($p<0.0001$, Figure 2).



There is no interaction effect; there is a cumulative increase in time with both physicians and nurses considered. The kernel density graphs indicate that the ER can become more efficient. However, the achievement of increased efficiency will require a change in healthcare provider behavior.

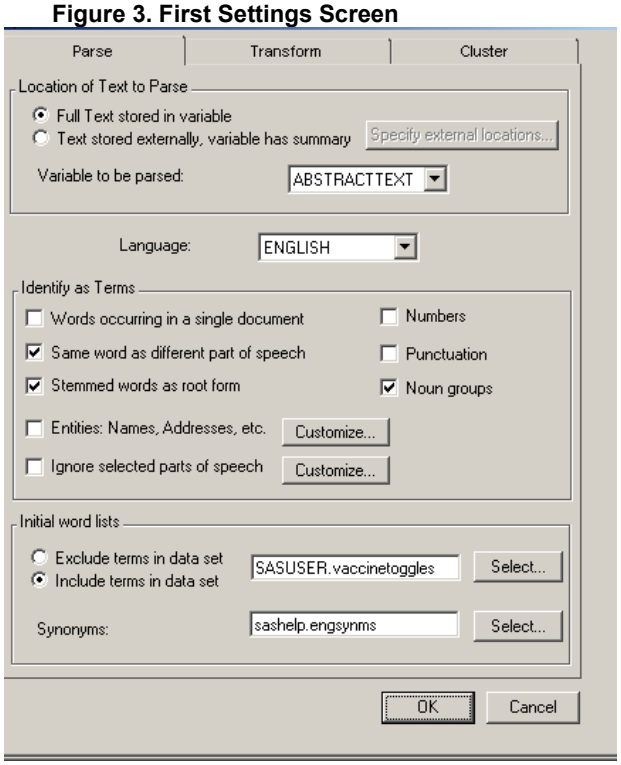
EXAMINATION OF PATIENT SEVERITY USING SAS TEXT MINER

In medical treatment, many of the documents on patient care exists in the form of chart notes and paper records. Currently, for billing and insurance requirements, hospitals and physicians must pay for manual extraction of information from those chart notes. This is a very time-intensive and costly process where text analysis can be used to reduce the cost of transferring information.

Another set of documents that are often collected but rarely analyzed are comments provided as part of a collection of survey data. Although the comments are read, and sometimes manually coded into categories, text analysis will make this information much more meaningful than it has been in the past. Once the documents are contained within a SAS dataset, and analyzed, it is possible to use more standard statistics and data mining techniques to further investigate them.

The Text Miner Node has three settings screens to examine. The first screen is given in Figure 3.

Figure 3. First Settings Screen

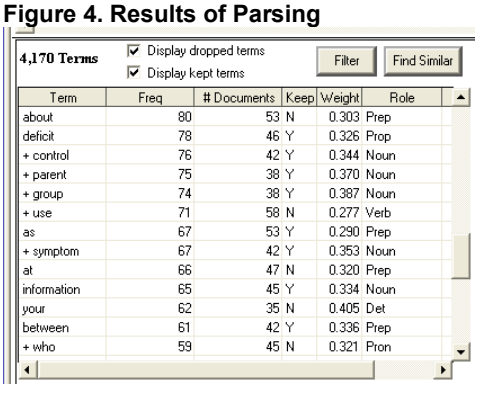


There are a number of defaults to consider. A standard ìstoplistî dataset will remove common words such as ìandî and ìtheî from consideration. The user can add words to the stoplist as needed, or create his own list. A second default is to exclude consideration of words that only occur in one document 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 possible to restrict attention to some specific terms by listing them in a dataset. Text Miner will only list terms from the specified dataset. The purpose of this step is to ëparseí the documents. Text parsing is a very technical process that is used to reduce the size of the documents to a manageable number. It also means that the software attempts to use grammar context to identify a specific part of speech for each term used. Modifiers are often connected to nouns to define ënoun groupsí (Figure 4).

Figure 4. Results of Parsing

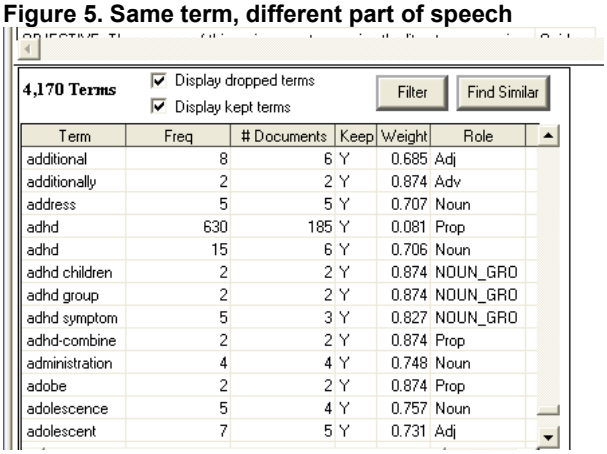


Term	Freq	# Documents	Keep	Weight	Role
about	80	53	N	0.303	Prep
deficit	78	46	Y	0.326	Prop
+ control	76	42	Y	0.344	Noun
+ parent	75	38	Y	0.370	Noun
+ group	74	38	Y	0.387	Noun
+ use	71	58	N	0.277	Verb
as	67	53	Y	0.290	Prep
+ symptom	67	42	Y	0.353	Noun
at	66	47	N	0.320	Prep
information	65	45	Y	0.334	Noun
your	62	35	N	0.405	Det
between	61	42	Y	0.336	Prep
+ who	59	45	N	0.321	Pron

The + signs indicate that there is more than one word connected to the phrase. Clicking on the ëTermí box will put the words in alphabetical order.

Notice that some of the terms have a ëYí or an ëNí. Any value with an ëNí is contained within the ëstoplistí file and is not used in the analysis. By unchecking the box, ëDisplay dropped termsí, all values with an ëNí are removed from the window; unchecking ëDisplay kept termsí removes all words with a ëYí. Consider Figure 5.

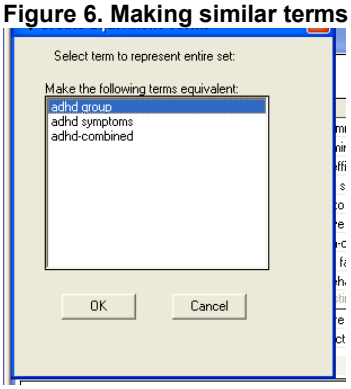
Figure 5. Same term, different part of speech



Term	Freq	# Documents	Keep	Weight	Role
additional	8	6	Y	0.685	Adj
additionally	2	2	Y	0.874	Adv
address	5	5	Y	0.707	Noun
adhd	630	185	Y	0.081	Prop
adhd	15	6	Y	0.706	Noun
adhd children	2	2	Y	0.874	NOUN_GRO
adhd group	2	2	Y	0.874	NOUN_GRO
adhd symptom	5	3	Y	0.827	NOUN_GRO
adhd-combine	2	2	Y	0.874	Prop
administration	4	4	Y	0.748	Noun
adobe	2	2	Y	0.874	Prop
adolescence	5	4	Y	0.757	Noun
adolescent	7	5	Y	0.731	Adj

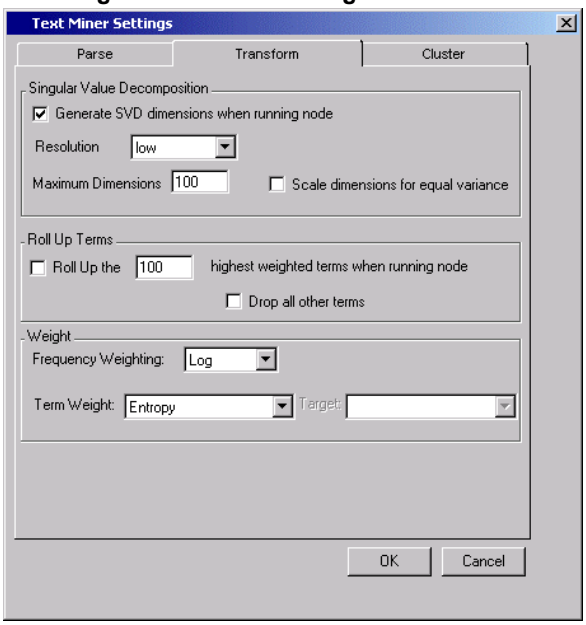
Ther term, éADHDí is both a proposition and a noun, as well as contained within a noun group. Text Miner will not allow these terms to be made equivalent, although the noun groups can be (Figure 6).

Figure 6. Making similar terms equivalent



However, an attempt to make éADHDí equivalent across different parts of speech will result in an error. This particular example generated 4,170 different terms. By unchecking the default éSame word as different part of speechí, reduces the number of terms to 3,914. Now it is possible to combine all references to éADHDí since parts of speech are no longer considered different.

Figure 7. Second Settings Screen



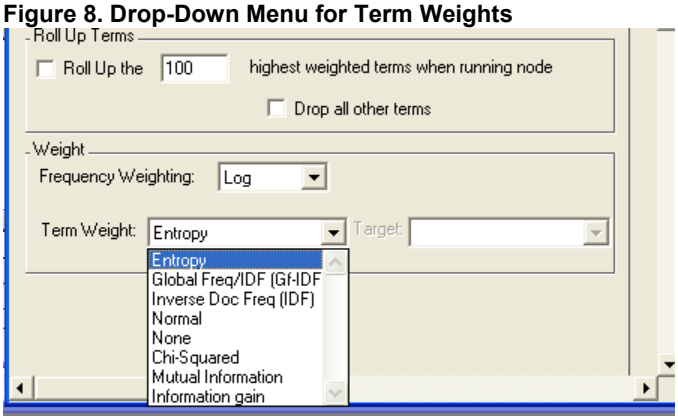
The second screen allows for the user to determine the method of reducing the wordlist matrix to a manageable size. The default is to use singular value decomposition. There are also several possible methods to weight the value of each term in the documents.

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.

Singular Value Composition defines a matrix of words by documents. The maximum dimensions (by default 100) box limits the size of this matrix. It can be increased depending upon the number of documents. However, the larger the matrix, the more time-consuming this process. The roll-up terms limits the wordlist to the top (100) highest weighted terms. A drop down menu will allow the user to change the weights (Figure 8.).

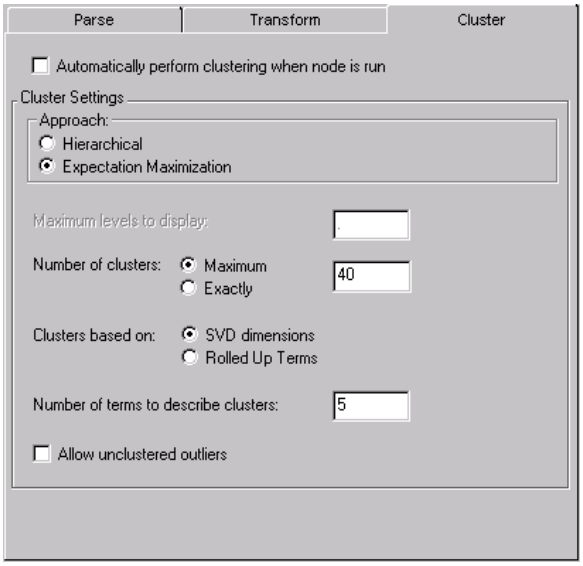
Figure 8. Drop-Down Menu for Term Weights



Entropy is the default. Terms that appear more frequently will be weighted lower compared to terms that appear less frequently. As such, it gives a result that is similar to Inverse Doc Freq. The last three assume the existence of a target variable in the dataset and cannot be used if there is no target.

Briefly, a status screen pops up, indicating that the singular value decomposition is being performed. The user can close this screen since the process will continue to run.

Figure 9. 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.

There is no one correct outcome to clustering text documents. Therefore, the user is free to change the settings in this and all previous boxes to get a desirable result.

It is possible that one or two physicians have a higher proportion of more critical patients. Therefore, it becomes necessary to define a ranking of initial patient complaints. Since these complaints are descriptive in nature, text analysis was used to divide the patients into a total of five categories (Table 1).

Table 1. Text Clusters of Patient Complaints

Cluster Number	Descriptive Terms	Cluster Label	Number of Patients
1	inhalation, smoke	Injury from smoke	161
2	vomiting, nausea, diarrhea, congestion, abd, allergic, sigu, migraine, reaction, n/v, accident, allergic reaction, car accident, neck pain, + stone, dizzy, toothache, severe headache, + need, + hive	Pain and nausea	240
3	flu, + symptom, like, sugar, flu symptoms, possible strep, blood, coughing, up, urine, strep, + low	Infection	31
4	chest, abdominal pain, chest pains, pressure, burning, abdominal pains, abdominal, + pain, chest pain, soa, urination, with, cough, + day, + day, back pain, in, right, back	Chest pain and potential heart problem	89
5	right, on, left, + fall, + shoulder, + knee, lt, arm, + eye, + leg, ankle, throat, lac, + foot, + swell, right, finger, + hurt, down, fall, broken, work, +	Injury	480

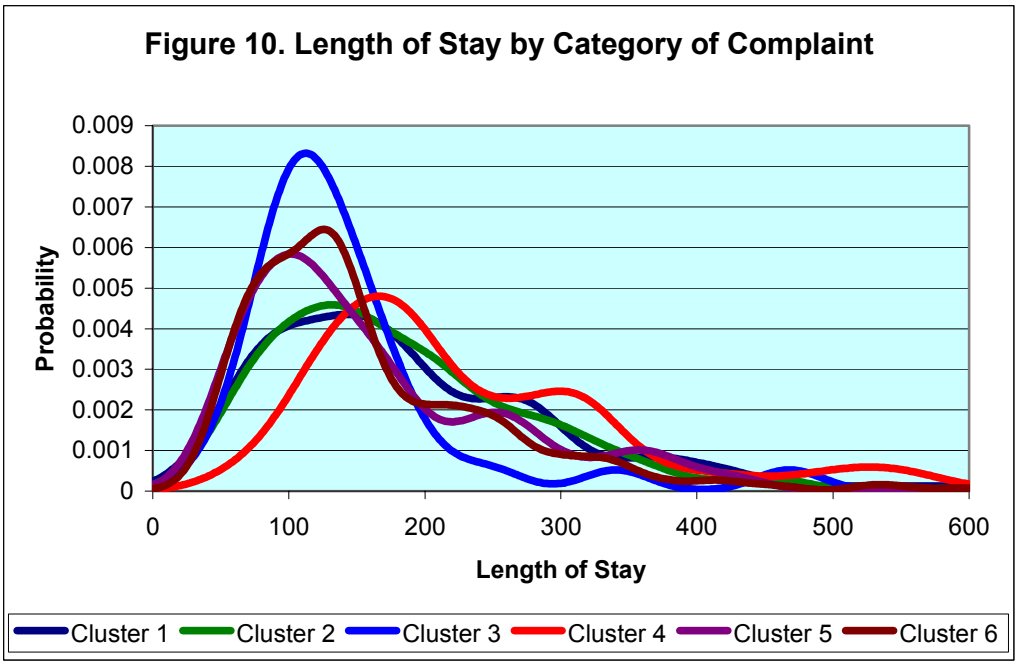
Cluster Number	Descriptive Terms	Cluster Label	Number of Patients
	injure, low, rt., injured, thumb, index		

Of the 43 patients admitted, fully 62% (27) are from categories 4 and 5; similarly 57% (22) of the transfer patients are from the same two categories. Of the patients in categories 1,2,and 3; 87%, 94%, and 87% respectively are discharged home compared to 77% from category 4 (and 91% from category 5). These differences are statistically significant (p=0.0004). It is clear that category 4 contains more critical patients. Since category 5 contains a high number of those admitted, but also has a high proportion discharged home, it is clear that it contains both critical and non-critical patients. Therefore, it was further sub-divided into categories 6,7, and 8 (Table 2).

Table 2. Modification of Injury Cluster

Cluster Number	Descriptive Terms	Cluster Label	Freq
5	shortness, head, cold, breath, left shoulder pain, lac, index finger, lip, + shoulder, lt, left, + pain	Minor injury	24
6	right, in, + knee, arm, + eye, + injury, + leg, ankle, throat, + foot, + hand, wrist, + swell, sore, right, rt, + leave, finger, back, + hurt, down, fall, work, + injure, + hurt, rt., + headache	More critical injury	362
7	mva, soa, ruq, rib pain, migrain, s/p, rapid, dysuria, test, epistaxis, nosebleed, back, + pain, + tooth, back pain, low, rt, + rib, + low, neck, severe, + day, mva, pt, flank pain, mid, + break	Minor injury	94

Category 6 contained 69% of the injured patients who were admitted or transferred. Since categories 5 and 7 have virtually the same rate of admission or transfer, they were then combined into one category (5). Figure 10 gives the length of stay by category.



Note that the most critical patients in cluster 4 have longer lengths of stay compared to the other clusters; category 3 has the shortest length of stay overall. The two categories of injury, 5 and 6 have somewhat shorter stays compared to patients in categories 1 and 2. Category 3 has the least amount of variability. The difference is statistically significant (p<0.0001). Category was compared by MD and by RN. The differences were not statistically significant so that the MDs are not handling different categories of patients. The average length of stay by category is given in Table 3.

Table 3. Length of Stay by Category of Patient Complaint

Category	Average LOS	Standard Deviation
Injury from smoke	190.22	111.02
Pain and nausea	190.98	117.65
Infection	142.58	82.79
Chest pain and potential heart problem	236.15	123.26
Minor injury	163.92	94.70
More critical injury	158.13	93.38

A second text analysis was used to examine patient charges. For the 1000 patients listed, approximately 30,000 charges were recorded. Ten clusters were identified as listed in Table 4.

Table 4. Text Clusters of Charges

Cluster Number	Descriptive Terms	Freq	Label
1	+ low, + instruction, dc, th/pro/, intravenous, injectons, intravenous th/pro/, cbc, panel, insertion, lock, hep/saline, hep/saline lock insertion, p/visit, chemo, not, thpy, infus, chemo p/visit	12370	IV charges
2	+ monitor, cardiac, cardiac monitoring, pulse, oxymetry, daily, ox, ox monitoring, hour, holter, + transport, rn, simple	764	Heart monitoring
3	+ dress, complex, simple, triage, triage complex, nursing re-assessment, + nurse, re-assessment	3703	Bandaging
4	ct, no, only, spine, wo, ct-head, dx, cervic-ap/lateral, ct-pelvic, ct-abd, lumbar-ap/lateral, thoracic-xray, regular, ct-chest, e.r., ct-spine, lamp, slit, abg, sacrum-xray, contrast, cervical-wo	1442	More complex X-ray and test
5	+ 4, iv, + 3, count, cell, glucose, fingerstick, + 1, + attempt, + tube, + 2, mini-neb, >, + 5, + 6, w/diff, hold, atrovent, protein/glucose, neb, alb/atrovent, csf, albuterol	1080	Diabetes and asthma
6	+ bone, xray-routine, + view, abd, xray, xray-portable, shoulder, + strap, acute, acute series-3, regular-ct, hip, w/pa, unilateral, rib, unilateral w/pa chest, xray-unilateral-two, xray-unilateral-tw	1308	X-ray series
7	oral, specimen, specimen collection, administration, med, collection, oxygen, exam, + point, pelvic exam assist, assist, admin, topical/rectal, topical/rectal med, + eye, ear, irrigation, abd/vaginal	3655	Urinalysis and Pelvic
8	urine, clean, clean catch, urinalysis, + wind, cleansing/irrigation, culture, pt, ptt, catch, catherization, cath, straight, foley, urine culture, stool, uretheral, magnesium, smear, wbcs, uric acid	1454	Bladder and kidney
9	extremity, troponin, test, qualitative, lipase, pregnancy, pregnancy test, screen, serum, amylase, rapid, + need, strep, rapid strep, prep, special, special needs, xray-ankle, xray-foot, + train	3894	Lab tests
10	im, injection, sq, ther/pro/d, antibiotic	1026	Antibiotic

The severity of the patientís condition was ranked by a healthcare provider independently of this analysis with 1 as the most severe and 10 as the least (Table 5).

Table 5. Severity of Charges by Complaint Category

Charge Cluster	Severity Ranking
IV charges	2
Heart monitoring	3
Bandaging	10
More complex X-ray and test	1
Diabetes and asthma	4
X-ray series	5
Urinalysis and Pelvic	9
Bladder and kidney	8

Bandaging is the least severe, with the least time involved in the interventions; complex testing and x-rays are the most severe with IV charges and heart monitoring close behind.

Charge Cluster	Severity Ranking
Lab tests	7
Antibiotic	6

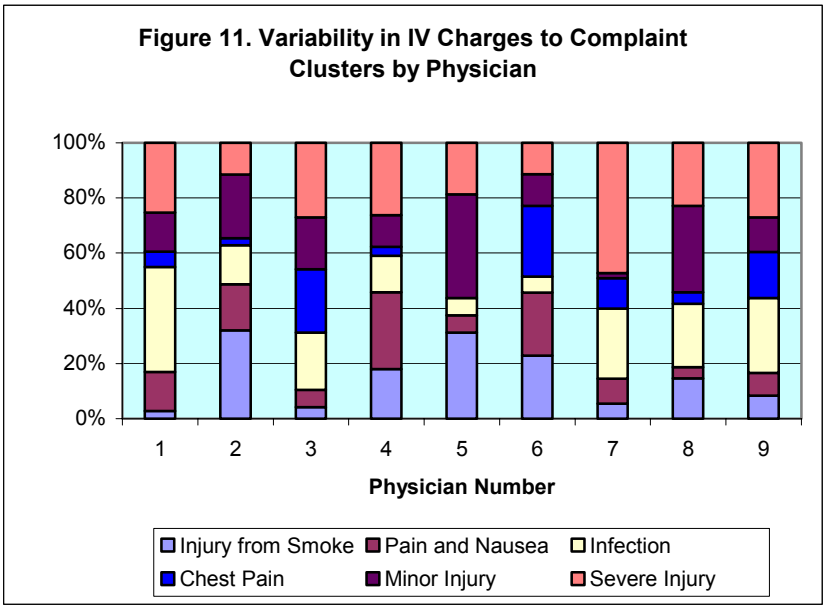
Reducing the patient disposition to admission, elope, and home care (due to empty cells), the results are statistically significant. Approximately 22% of the patients with charges in cluster 2 (heart monitoring) are admitted to a hospital setting. As a close second, patients in category 8 are admitted 19% of the time while patients in cluster 5 are admitted 10% of the time. The difference is statistically significant with p=0.0376. While patients in cluster 4 require the most time, only 5% are admitted. The relationship of charges to complaints is given in Table 6.

Table 6. Charge Clusters by Patient Complaint

Charge Cluster/Complaint Cluster	Injury from Smoke	Pain and Nausea	Infection	Chest Pain	Minor Injury	Severe Injury
IV charges	1229 (9.9%)	1684 (13.8%)	2547 (20.6%)	1271 (10.3%)	2867 (23.2%)	2772 (22.4%)
Heart monitoring	125 (16.4%)	157 (20.6%)	45 (5.9%)	201 (26.3%)	62 (8.1%)	174 (22.8%)
Bandaging	343 (9.3%)	465 (13.1%)	558 (15.1%)	318 (8.6%)	1175 (31.7%)	824 (22.2%)
More complex X-ray and test	139 (9.6%)	145 (10.1%)	300 (20.8%)	212 (14.7%)	166 (11.5%)	480 (33.3%)
Diabetes and asthma	167 (15.5%)	218 (20.2%)	261 (23.2%)	128 (11.4%)	104 (9.6%)	212 (19.6%)
X-ray series	104 (8.0%)	136 (10.4%)	159 (12.2%)	187 (12.8%)	413 (31.6%)	329 (25.2%)
Urinalysis and Pelvic	372 (10.2%)	582 (15.9%)	741 (20.3%)	517 (14.2%)	523 (14.3%)	920 (25.2%)
Bladder and kidney	119 (8,2%)	233 (16,0%)	328 (22.7%)	114 (7.8%)	300 (20.6%)	360 (24.8%)
Lab tests	297 (7.6%)	496 (12.7%)	718 (18.4%)	395 (10.1%)	1143 (29.4%)	846 (21.7%)
Antibiotic	156 (16.2%)	90 (8.8%)	156 (16.2%)	111 (10.8%)	247 (24.1%)	266 (25.9%)

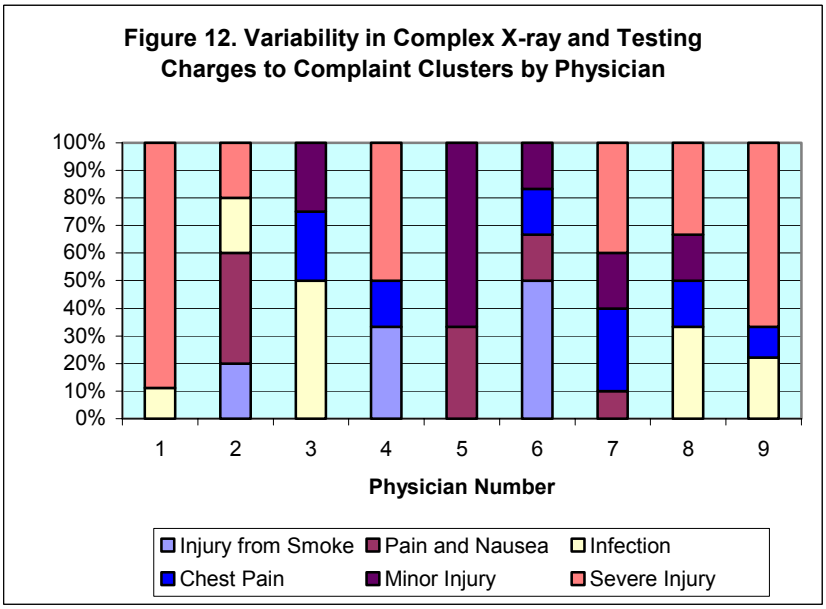
As expected, the infection group has a large percentage of antibiotic use, combined with an almost similar proportion of lab use (to discover the nature of the infection). Surprisingly, the injury groups have an even higher proportion of antibiotic use. As expected, almost 1/3 of the complex x-ray group come from the patients with severe injuries; 1/3 of the bandaging is for the minor injury group and 1/4 of the heart monitoring is for the group with chest pains. Asthma and diabetes testing is almost uniform across the categories of complaints.

It is of interest to note that there is considerable variability in the way physicians prescribe to the six different complaint clusters (Figure 11).



In particular, note that physician 7 has almost no IV charges for minor injury whereas physician 5 has a considerable proportion. The result is statistically significant (p=0.0048). Conversely, Physicians 4 and 5 have almost no IV charges for patients with infection but physician 7 as a considerable proportion.

Similarly, Figure 12 is a graph of complex x-ray and testing by physician.



This result is also statistically significant (p<0.0001), using Fisherís Exact Test since there are a number of empty cells. Physician 5 uses complex x-ray for minor injury and pain while physician 9 reserves these tests for severe injury, as does physician 1.

DATA VISUALIZATION USING LINK ANALYSIS

Another way of investigating the relationship of complaints to charges is through link analysis (Figure 13). The initial output screen gives the chi-square distribution for the input variables. The use can request through the View Sub-Menu, detailed settings. The result is given in Figure 14.

Figure 13. Initial screen for Link Analysis Node

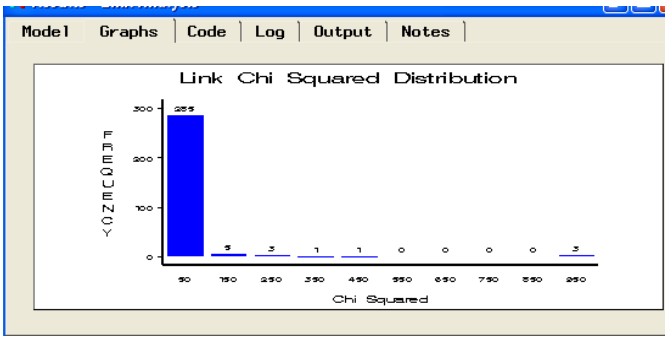
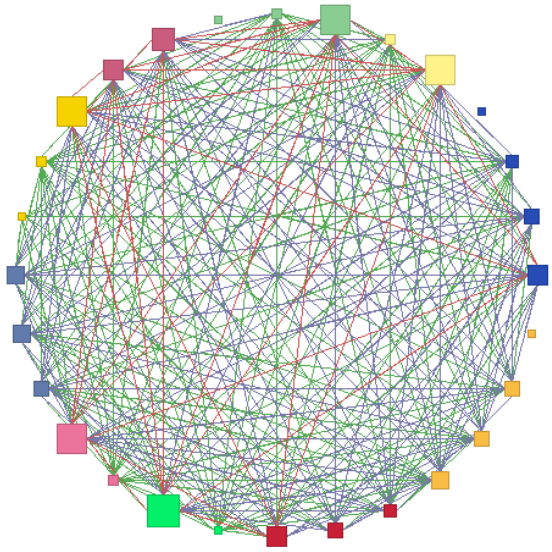


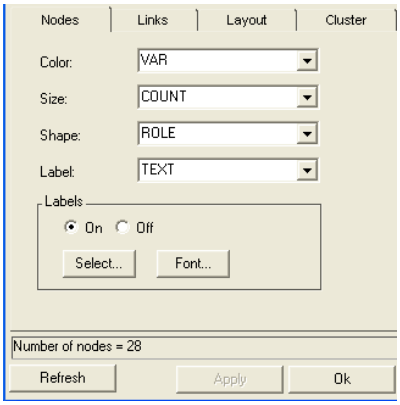
Figure 14. Detailed Graphs in Link Analysis Results



The size and color of the nodes give the extent of the correlations. Red denotes a higher level of correlation; larger nodes denote more values in that category.

The user can go to Action>display to modify the graph. Figure 15 indicates how the labels can be turned on.

Figure 15. Display Control in the Link Node



The user can also change the orientation of the graphic to a variety of different types (Figure 16) by pressing the layout button in the Display Control. Figure 17 gives the MDS graphic.

Figure 16. Potential Layouts of Link Analysis

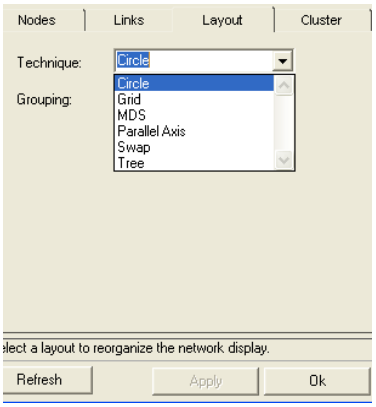


Figure 17. MDS Graphic Display

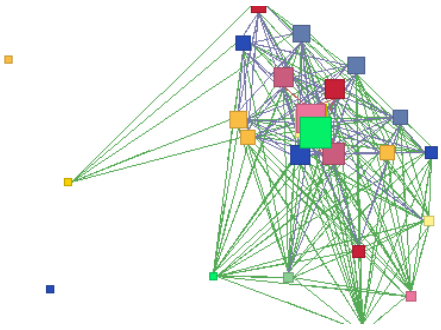
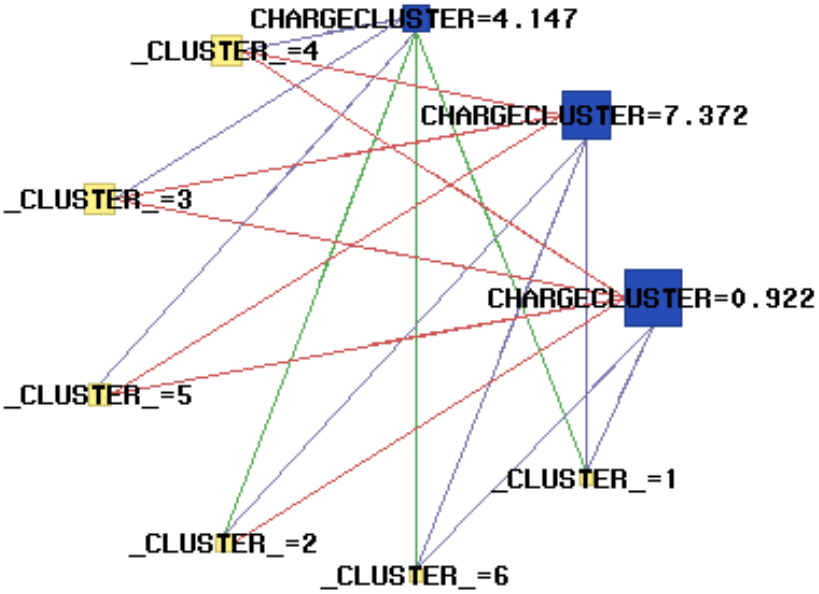


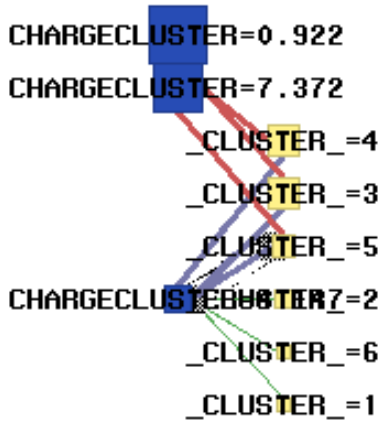
Figure 18 gives the links to relate patient complaints to charges in the ER.

Figure 18. Link Analysis of Charges to Complaints.



Link analysis is a pictorial representation of correlations between variables. In the circle graph depicted in Figure 11, the strength of the correlation is represented by the color, with red as the highest correlation and green as the lowest. Patient complaints 1-6 are represented on the left-hand side of the figure, with charge clusters on the right. The size of the node represents the size of the cluster relationship. Charge cluster number 1 is the largest size since it has such a large number of items contained within. It has the highest correlation with complaint clusters 2,3,4,5 and weaker correlation with clusters 1 and 6. Another way of examining the link analysis is as a tree (Figure 19).

Figure 19. Tree View of Link Analysis



TEXT ANALYSIS TO EXAMINE HOSPITAL BILLING AND QUALITY RANKINGS

The purpose of this section is to examine the feasibility of using text analysis to investigate coded information. Although codes do not have regular grammar and syntax, each code, representing a discrete category, can be examined using alphabetic notation. In this example, ICD-9 codes that are used to represent patient diagnostic information were examined. ICD-9 codes are 5-digits, of which the first three represent a general patient category while the last 2 digits are more specific. ICD-9 codes are routinely used in the health profession for billing purposes. Because different hospitals do not uniformly record ICD-9 codes, there is a lot of statistical noise in the system. It is the nature of text analysis to examine unstructured text, and the method allows for a significant amount of noise.

ICD-9 codes were developed by the World Health Organization as shorthand for patient illness, complication, and disease. ICD-9 codes that are similar in definition are also similar in number. The ICD-9 code can have 5 digits, with the first 3 digits defining the basic disease. For example, é250í codes diabetes. If the last digit is é1í, (25001) the condition is Type I (insulin-dependent, juvenile diabetes); a é2í (25002)in the last digit signifies Type II (non-insulin, adult onset diabetes). The 4th digit is used to signify different diabetic complications. For example, 25041 indicates a Type I diabetic with renal failure. ICD-9 codes represent a coded language that can be investigated using text analysis, and cluster analysis (or classification) using a text format.

The ordering of the first three numbers is arbitrary. The next lowest code to diabetes is é246í, which deals with thyroid disorders and is not related to diabetes. The next highest code is é251í for hypoglycemic coma, and does have some relationship to diabetes. However, code é252í deals with diseases of the parathyroid gland. The first three numbers represents a basic category of patient illness. There remain thousands of such codes. It is possible to treat the codes as text rather than as numbers and to use the numbers as letters to relate codes in clusters.

These codes can be examined to determine whether the data have been recorded uniformly, and whether various hospitals use similar definitions. Ultimately, the information in the ICD-9 codes submitted with billing data must be compared to information in patient charts. Currently, most hospitals rely on manual extraction of information from patient records, requiring many extractors. With hundreds of pages to examine, manual extraction can result in missed data and missed diagnosis codes. Therefore, it will be cost-effective and more accurate to extract information automatically from patient charts, which may lead to higher numbers of risks, and can be used for consistent reporting. However, statistical methods are not really applicable to extraction of textual information.

The standard procedure used by insurance providers, industry watchdog groups, and professional societies to examine hospital quality and cost effectiveness has been an equation of the form:

$$y = B_0 + B_1x_1 + \ldots + B_kx_k + e$$

where y is the outcome variable, $x_1, x_2, \ldots, x_k$ are the variables used to predict the value of y, e is a random error factor, and B_i determines the contribution of the variables x_i. The variable y can be continuous (such as length of stay or costs), or discrete (such as mortality or complication). Each x variable denotes the presence or absence of a risk factor for a particular patient. The value 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 value of y is the linear combination of weights. If y is a discrete variable such as mortality, an optimum threshold value of found, and mortality is predicted if the sum of the weights exceeds that threshold.

To determine a ranking, the expected value of y is compared to the actual value of y, and the difference is computed. If this difference is positive then the overall patient outcomes were better than expected and the higher the healthcare facility will be ranked. The opposite is true if the expected difference is negative; the patient outcomes were lower than expected, and the healthcare facility will be ranked lower. At this point, hospital administrators can gauge how their facilities are performing compared to other facilities through publicly available quality rankings that are based on publicly available Medicare billing data; insurance companies and national societies include clinical information.

Because there are so many possible risk factors that can be used to develop an estimated ranking ñ specifically, the entire set of ICD-9 codes ñ many are traditionally eliminated at the outset because they lack statistical significance in a pairwise comparison. Given a sufficiently large number of patients, almost any factor can become significant. Analysts traditionally use a stepwise procedure so that the most important risk contributors are included in the model. This traditional way of analyzing quality can result in many different models, and each organization that has developed a model can use one with its own customized modifications. The models are rarely validated; reliability is not usually demonstrated. The process, and the model equation are not usually made available. If y is discrete, then a measure of the accuracy of the model is in the c statistic; if y is continuous, the measure is r^2. The two measures are generally relatively low in models used for predictions, but they, too, are rarely provided. Therefore, much of the difference in predicted quality remains unaccounted for in the models.

In order to create these models, analysts assume that patient risk factors are uniformly entered by all providers. We know this canít possibly be true. If one hospital regularly under-reports the risk factors facing its patients, then the expected patient outcomes will have fewer weights compared to other hospitals, and the difference between expected and actual patient outcomes will be lower compared to other hospitals, resulting in a low ranking. Therefore, the process rewards hospitals that tend to over-report on risk factors.

For example, there are 51 ICD-9 codes related to diabetes. Consider just the following five:

1. 250 Diabetes Mellitus without mention of complications
2. 25000 Type II diabetes mellitus without mention of complications
3. 25001 Type I diabetes mellitus without mention of complications
4. 25002 Type II diabetes mellitus without mention of complications uncontrolled
5. 25003 Type I diabetes mellitus without mention of complications uncontrolled

A physician has sole discretion in documenting ëuncontrolled diabetesí with very few guidelines:[3, 4]

1. Symptomatic hyperglycemia
2. Fasting blood glucose level above 300 mg/dl
3. Hgb A1-C two times the usual limit of normal
4. Frequent swings between hyperglycemia and hypoglycemia

If a physician is more in the habit of adjusting insulin treatment without documenting ëuncontrolledí then one hospital will have fewer patients with a more severe condition than another hospital where the physician continually documents ëuncontrolledí after one measurement above 300.

A novel approach to analyzing ICD-9 codes is to treat them as text rather than as categories. In that way, similarities between codes can be related to similarities in patient conditions, taking full advantage of the stemming properties contained within the codes. Similar stemming occurs in medication names so that patient information can be clustered.

Instead of relying on ICD-9 codes that require the invalid assumption that manual extraction across hospitals is consistent, we propose an alternative method for identifying patient severity was also developed by using the medications in the pharmacy database (Pyxis, Millersville, MD). The first challenge of the data analyst is to audit and to clean the data needed to perform the analysis. First, the data must be obtained from some data source(s). Since more than one data source is involved, it is necessary to integrate the data, which involves a laborious process of determining when data are compatible: whether formats are uniform or not and whether units of measurement are the same or not. Even when data come from one source only, there is still some cleansing needed. In the particular case of text mining, information may be pulled from text fields in records, written transcripts, or even free-format documents (such as emails).

Once the data are obtained, further processing is needed. Since the data are not completely clean, nor in the right format, some form of transformation must be applied to it. For text mining, some of the preprocessing involves fixing typos, grammatical errors, getting rid of erroneous characters, and even substitution of some terms by synonyms in order to homogenize the data. It is also normal to apply some syntactic analysis (like stemming) at this point.

Stemming means finding and returning the root form (or base form) of a word. Stemming enables the investigator to work with linguistic forms that are more abstract than those of the original text. For example, the stem of **grind**, **grinds**, **grinding**, and **ground** is **grind**. The document collection often contains terms that do have the same base form but share the same meaning in context. For example, the words **teach**, **instruct**, **educate**, and **train** do not have a common stem,

but share the same meaning of **teach** Text mining can relate words with similar stems. The capability can be extended to numeric codes. Grammar and syntax are not important in examining the ICD-9 codes using text mining. However, stemming has value since the ICD-9 codes have base numbers. For example, é250í represents diabetes. From that base, é25001í, é25002í represent additional diabetic conditions. é25001í represents uncomplicated diabetes but é25040í represents a diabetic patient with renal problems.

An experimental analysis was performed to examine formulas used to predict patient risks to examine the feasibility of information extraction from text sources, and to investigate the question of uniform data entry across hospitals. ICD-9 codes have been analyzed as unstructured text to define patient profiles. Up to nine secondary diagnoses can be listed in the Medicare reporting form (UB-92). Some of the secondary diagnoses are for complications from treatment, but many reflect co-morbid illnesses. A study was conducted using data provided by the Kentucky Hospital Association for all hospitals in Kentucky performing cardiovascular procedures. There were a total of over 16.000 cardiovascular procedures available for examination. Table 1 contains the defined patient profiles.

Table 7. Patient Profiles Developed From Text Mining of Secondary Diagnosis Codes

No.	Code	Description	No.	Code	Description
		General Risk Factors			**Severe Cardiac Risks**
1	2449	Acquired hypothyroidism	**5**	4148	Heart ischemia
1	25001	Diabetes, uncomplicated	**5**	4240	Endocardium disease
1	25002	Diabetes, uncomplicated	**5**	4271	Supraventricular paroxysmal tachycardia
1	2768	Fluid disorder, electrolyte and acid-base balance	**5**	42741	Ventricular fibrillation and flutter
1	27801	Obesity, other hyperalimentation	**5**	3970	Rheumatic fever, other endocardial structures
1	41402	Coronary atherosclerosis	**5**	49121	Chronic obstructive bronchitis
1	5939	Unspecified kidney and ureter disease	**5**	4275	Cardiac arrest
1	V4581	Aortocoronary bypass	**5**	49120	Chronic obstructive bronchitis
1	7140	Rheumatoid arthritis	**5**	4589	Unspecified hypotension
1	490	Bronchitis	**5**	4109	Acute myocardial infarction
		General Vascular Risk Factors			**Severe Respiratory Risks**
2	4264	Disease, heart conduction	**6**	486	Pneumonia
2	42789	Cardiac dysrhythmia	**6**	5121	Iatrogenic pneumothorax
2	43310	Occlusion & stenosis, carotid artery, cerebral infarction	**6**	51881	Acute respiratory failure
2	4401	Renal artery arteriosclerosis	**6**	9973	Surgical respiratory complication
2	4439	Peripheral vascular disease	**6**	41400	Coronary atherosclerosis
2	5533	Diaphragmatic hernia	**6**	99889	Other specified complication not classified elsewhere
2	71590	Osteoarthrosis	**6**	49121	Chronic obstructive bronchitis
2	78057	Sleep apnea	**6**	4928	Emphysema
2	V1259	History, certain other diseases	**6**	389	Hearing loss
2	V173	Family history, certain chronic disabling diseases	**6**	2639	Protein-calorie malnutrition
		Moderate Cardiac Risks			**Severe Renal Risks**
3	2761	Fluid disorder, electrolyte and acid-base balance	**7**	25040	Diabetes, renal manifestations
3	2766	Fluid disorder, electrolyte and acid-base balance	**7**	2767	Fluid disorder, electrolyte and acid-base balance
3	2875	Unspecified thrombocytopenia	**7**	40391	Hypertensive renal with renal failure disease
3	42732	Atrial fibrillation and flutter	**7**	58381	Nephritis and nephropathy
3	4294	Functional heart disturbance, long-term effect	**7**	5849	Acute renal failure
3	99811	Hemorrhage, hematoma, seroma	**7**	5990	Urinary tract infection
3	99812	Hemorrhage, hematoma, seroma	**7**	9975	Surgical Urinary complication
3	4582	Iatrogenic hypotension	**7**	2765	Volume depletion
3	4239	Hemopericardium	**7**	25041	Diabetes, renal manifestations
3	7806	Fever	**7**	585	Chronic renal failure
		Complicated Diabetes			**Severe neurological risks and complications**
4	25051	Diabetes, ophthalmic manifestations	**8**	2762	Fluid disorder, electrolyte and acid-base balance

No.	Code	Description	No.	Code	Description
4	25060	Diabetes, neurological manifestations	8	43491	Cerebral artery occlusion, with cerebral infarction
4	25061	Diabetes, neurological manifestations	8	436	Acute Cerebrovascular disease
4	25062	Diabetes, neurological manifestations	8	5185	Pulmonary insufficiency following trauma and surgery
4	3371	Autonomic nervous system disorder	8	5601	Paralytic ileus
4	3572	Polyneuropathy, in diabetes	8	78551	Cardiogenic shock
4	36201	Retinopathy, diabetic	8	99702	Central nervous system complication
4	25052	Diabetes, ophthalmic manifestations	8	99859	Postoperative infection
4	25050	Diabetes, ophthalmic manifestations	8	9974	Digestive system complication
4	4231	Pericardium, other disease	8	5845	Acute renal failure with tubular lesion of necrosis

The UB-92 form used for Medicare billing has nine column fields to record information by ICD-9 codes. Text analysis can only examine one field at a time. For analysis purposes, the nine fields were combined into one, with the nine ICD-9 codes for an individual patient defining one text string. Then a singular value decomposition was performed, which creates a matrix where ICD-9 codes that appear in combination can be identified, based on projecting patient profiles into a multidimensional space. An Expectation-Maximization clustering algorithm was then run to cluster the patient profiles (as discussed in the methods section). SAS Enterprise Miner for text was used to perform the analysis. Table 1 gives the results of the text analysis. A total of 8 clusters were identified in the analysis.

For example, cluster one contains uncomplicated diabetes (both Type I and Type II) as well as other problems such as arthritis and hypothyroidism. The majority of patients in this cluster are given a risk factor of 1 or 2 (low risk) rather than 3 or 4 (high risk). In contrast, cluster 4 is primarily focused on diabetes with complications, mostly retinopathy or neuropathy. Both conditions indicate that the patient has prolonged, probably uncontrolled diabetes. Patients in this cluster generally are assigned a risk factor of 4. The difference between recording 25001 and 25003 is slight. Yet, the results could be that a hospital with a high proportion in cluster 1 will be ranked low on the standardized risk adjusted regression equations. A hospital with a greater proportion of patients in cluster 4 compared to other hospitals is better at documenting serious complications from diabetes than hospitals with a higher proportion in cluster 1.

Similarly, cluster 2 identifies patients generally who have a need for open heart surgery. They represent modest risk factors. Hospitals with a high concentration of patients in cluster 2 is probably under-reporting on risk factors, and will end up ranked lower based upon regression equations. Cluster 7 is associated with patients with critical risks, primarily dealing with kidney failure or even dialysis. Note also that bronchitis, ë490í is represented in the mild cluster 1 while ë49121í or chronic bronchitis is in clusters 5 and 6.

Different hospitals have different proportions of patients in the above clusters (Table 8).

Table 8. Proportions of patients in each cluster by hospital

Hospital	1	2	3	4	5	6	7	8
1	35	9	23	1	8	6	11	8
2	40	8	20	3	9	9	7	5
3	35	10	21	1	12	4	7	9
4	32	10	26	1	9	8	7	8
5	25	13	34	3	11	7	4	9
6	14	14	36	8	3	8	11	6

In addition, categories of codes can be flagged, and examination by specific diagnoses. For example, the group ë250í signifies diabetes. When only patients with a diagnosis were flagged, 6 clusters emerged (Table 9).

Table 9. Clusters for Patients With Diabetes (3864 Patients)

Codes in Cluster	Diagnoses
25093, 25083	Unspecified complications, unspecified manifestations
25062, 25052, 25042, 25072, 25013	Neurological, Ophthalmic, Renal, Peripheral circulatory disorders, diabetes coma
25000	Diabetes
25001	Type I diabetes
25060, 25050, 25040, 25070, 25080	Neurological, Ophthalmic, Renal, Peripheral circulatory disorders
25002	Type II diabetes

Note: the last digit represents 1 for Type I, 2 for Type II, and III for not specified. The code is a zero is the diabetes type is not identified. It is clear from the specifics that clusters 1, 3, 4, and 5 have a lower ranking of severity compared to clusters 2 and 5. The same hospitals listed in Table 2 differ by as much as 15% the proportion of patients in those higher risk categories. A similar result occurs for patients with lung diseases (Table 10).

Table 10. Clusters for Patients with Lung Diseases (4565 Patients)

Codes in Cluster	Diagnoses
49300. 49400, 49110, 49100	Extrinsic asthma, Bronchlectasis, Chronic bronchitis-simple or mucoputulent
48210, 48282, 49000, 49600	Pneumonia from Psudomonas, Pneumonia from unspecified bacteria, Unspecified bronchitis, COPD
49120, 48220, 49600	Chronic obstructive bronchitis, Pneumonia from Hemophilus influenzae, COPD
48600, 49320, 49000, 49600	Pneumonia organism unspecified, Chronic obstructive asthma, Unspecified bronchitis, COPD
49390, 49190, 49600	Unspecified asthma, Chronic unspecified bronchitis, COPD
49280	Emphysema
49600	COPD
49121, 48239, 48090, 48241, 49600	Chronic obstructive bronchitis, Pneumonia from Hemophilus influenzae, Viral unspecified pneumonia, Pneumonia from staphylococcus, COPD

Note that 49600, chronic obstructive pulmonary disease (COPD), appears in every cluster except the first and sixth. Therefore, it has little value in discriminating in the level of severity.

CONCLUSION

Data mining of clinical data will significantly increase the amount of information concerning costs and quality in patient care. More information will permit hospitals to become more efficient, and an investigation in the variability of practice will enable hospitals to optimize decisions that impact care.

REFERENCES

1. Healthgrades. Healthgrades. *Healthgrades*. Available at: www.healthgrades.com. Accessed August 23, 2001, 2001.
2. Anonymous. *DRG Guidebook: A comprehensive reference to the DRG classification system*. 17th ed. Reston, Virginia: St. Anthony Publishing; 2001.
3. Anonymous. Coding & DRG Notes: Diabetes mellitus. Available at: http://www.medicarequality.org/PEPP/PDF/DRGNotesDmUncontrolled.pdf.
4. Johnson ML, gordon HS, Peterson NJ, et al. Effect of definition of mortality on hospital profiles. *Medical Care.* 2002;40(1):7-16.
5. Mosteller F, Wallace D. *Applied Bayesian and classical inference: the case of the Federalist Papers*. New York: Springer-Verlag; 1984.
6. Holmes D, Forsyth R. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing.* 1995;10(2):111-127.
7. Martens BVdV. IST 501: Research Techniques for Information Management. Available at: http://web.syr.edu/~bvmarten/index.html. Accessed 2002, 2002.

ACKNOWLEDGMENTS

The author would like to acknowledge support from NIH for this project # 1R15RR017285-01A1, *Data Mining to Enhance Medical Research of Clincal Data.*

CONTACT INFORMATION

Author Name	Patricia B. Cerrito
Company	University of Louisville and Jewish Hospital Center for Advanced Medicine
Address	Department of Mathematics
City state ZIP	Louisville, KY 40292
Work Phone:	502-852-682/502-560-8534
Fax:	502-852-7132
Email:	pcerrito@louisville.edu
Web:	www.math.louisville.edu