5-2006

Text mining comorbidity codes in the analysis of cardiopulmonary rehabilitation data.

Jennifer Ferrell 1982-
University of Louisville

TEXT MINING COMORBIDITY CODES IN THE ANALYSIS OF CARDIOPULMONARY REHABILITATION DATA

By

Jennifer Ferrell

A Thesis
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science in Public Health

School of Public Health
Department of Biostatistics

May 2006

TEXT MINING COMORBIDITY CODES IN THE ANALYSIS OF CARDIOPULMONARY REHABILITATION DATA

By

Jennifer Ferrell
University of Louisville

A Thesis Approved on

April 14, 2006

by the following Thesis Committee

____
Thesis Director

____

____

ACKNOWLEDGMENTS

I would like to acknowledge my thesis advisor, Dr. Caryn Thompson, for her patience and approval of my thesis. I would also like to thank the other committee members, Dr. Guy Brock and Dr. Patricia Cerrito, for their helpful comments and assistance. I would also like to thank my family and friends for their patience with me throughout this process.

ABSTRACT

TEXT MINING COMORBIDITY CODES IN THE ANALYSIS OF CARDIOPULMONARY REHABILITATION DATA

Jennifer Ferrell
Advisor: Dr. Caryn Thompson
May 13, 2006

The purpose of this paper is to examine the process of text mining and using the results to show the possible benefits of cardiopulmonary rehabilitation. The 555 patients enrolled in the study were receiving inpatient cardiopulmonary rehabilitation. Each patient had comorbidity codes associated with them. These codes are secondary diagnoses to the cardiac or pulmonary event that resulted in their hospitalization. The patients had secondary conditions ranging in number from 1 to 10. The patients were assessed at admission and discharge for functional independence. Since there are numerous comorbidity codes for each patient, it would be difficult to analyze each one. Therefore, we can text mine these codes to create meaningful clusters to help in the analysis. This paper explains the process and theory of text mining and clustering. We use these results to perform statistical analysis to examine the benefits of cardiopulmonary rehabilitation.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Currently, Medicare does not consider cardiopulmonary and other chronic illness patients as having a rehabilitation classification. This thesis focuses on data from a study, which was conducted to determine whether patients with chronic illnesses receive the same benefits from physical therapy as would persons with orthopedic problems.

The study, which was conducted at Frazier Rehabilitation Center, resulted in a data set of 555 records from patients with chronic illnesses related to cardiac and pulmonary problems. The patients were enrolled in an inpatient rehabilitation program. Each patient was assigned comorbidity codes when they entered the hospital. Comorbidity codes represent the conditions that exist along with the main cause of the patients' need for hospitalization. These codes are based on the ICD-9 codes that are generally used in hospitals for insurance billing.

Since most patients have several comorbidity codes, it is very difficult and time consuming to evaluate each patient and all of their separate codes. In this thesis, text mining, a relatively new technique, was used to create text strings for each patient in the data set described above, and to extract and compare the similarities among the patients. Text mining is slowly catching on in the medical informatics field. Currently, it is used, for example, to assist in the microarray analysis of genes Chen *et al* (2005). However, not much research has been conducted on applying text mining to numerical codes. Recently, Cerrito (2003) demonstrated the application of text mining to ICD-9 codes in order to demonstrate how the codes are used in insurance billing.

Text mining is used to process textual information so that it can be used in the data mining techniques such as clustering. Clustering is a multivariate technique used to find similarities between observations to group the observations together. Since the ICD-9 codes represent words and are otherwise nominal data, text mining was chosen over data mining in this study, following Cerrito *et al* (2003). The main difference between text mining and data mining is that data mining processes ordinal and interval data while text mining processes nominal data. Text mining and clustering are very effective for creating useful variables. In the health industry, clinical trials are sometimes kept smaller so that text analysis is not too time consuming. For example, if health professionals are analyzing surveys by hand, they may only be able to analyze 20 surveys since it is a time consuming process. By using a text miner computer software, thousands or more surveys can be analyzed rather quickly compared to analyzing surveys by hand. As the knowledge and literature of text mining increases, it may see greater application in health related studies.

In this thesis, clustering and text mining were used to group similar patients together for analysis purposes. Clustering and text mining were performed in SAS Enterprise Miner®. The resulting clusters were ranked based upon severity of comorbidity codes and included in predictive models. The purpose of this study was to assess therapeutic benefit, through exploration of relationships between severity and age.

The second chapter will present a basic background of cardiopulmonary rehabilitation, medical informatics, and text mining including ICD-9 codes. The third chapter will provide an overview of the data as well as the theory behind text mining and clustering. A brief example and a summary of the application of text mining to the dataset will be described. The fourth chapter will display the results and an in-depth discussion of how text mining and clustering were applied to the dataset. Lastly, statistical analysis and conclusions about the data will be discussed in Chapter 5. A discussion of the advantages and disadvantages of text mining is also included in the final chapter, along with suggestions for improved methodology and further research.

CHAPTER 2

BACKGROUND

2.1 Cardiopulmonary Rehabilitation

Cardiopulmonary rehabilitation is treatment administered to patients who have suffered a cardiac or pulmonary event in the hopes of restoring their physical state. Rehabilitation focuses on improving not only physical problems from the cardiopulmonary event, but assessing the risk factors that led to the incident. Each program is tailored to the individual’s needs, but a general format is followed. The program usually consists of three phases defined by Foley *et al* (1993): inpatient, immediate outpatient and community-based.

Inpatient rehabilitation focuses on getting the patient to perform daily activities such as bathing and using the restroom without assistance. This phase of rehabilitation (Phase 1) usually occurs when patients are still in the medical center and is overseen by the patient’s nurse or doctor. Immediate outpatient rehabilitation involves more of the changes in lifestyle such as reducing risk factors and increasing physical functionality. This occurs immediately after the patient is discharged from the hospital. Most patients in this second phase (Phase 2) are discharged to their homes and visit physical rehabilitation centers for treatment. The third phase (Phase 3) is a community-based rehabilitation. This is a continuation of the outpatient phase, but is left up to the patients to continue on their own. This involves patients consciously continuing exercises and reducing risk factors without the assistance of a physical therapist (Foley *et al*,1993).

Some consider Phase 1 to have two subphases: Phase 1A and Phase 1B. Phase 1A considers the rehabilitation of patients still in medical care. However, some patients are sent to nursing care facilities or inpatient rehabilitation centers. These patients are considered to be in Phase 1B. These patients are not well enough to be discharged home, but do not require acute medical care. Another difference between these subphases is the reimbursement process (Sansone *et al*, 2002). These subphases may provide better analysis of the benefits of rehabilitation. In general, Foley's three-phase program is more widely accepted.

In the past, most studies conducted on cardiopulmonary patients have only involved middle-aged men (Lear *et al*,2001). More recent studies have included men and women of all ages and races who have experienced a cardiopulmonary event. Rehabilitation usually happens shortly after the event has occurred. Generally, the sooner the rehabilitation program can begin, the faster the patient can get on the road to recovery. Also, as rehabilitation takes place, it is expected that the probability of having another cardiopulmonary event will decrease.

Part of cardiopulmonary rehabilitation is learning and reducing the risk factors that first caused the event. The combination of increasing physical well-being, mental well-being and learning about past health mistakes all contribute to the suggestion that cardiopulmonary rehabilitation can benefit patients.

2.2 FIM Scores

Functional Independence Measure (FIM) scores are measurements used to assess how well a patient is mentally and physically performing on their own. These scores range from “1” to “7” on each item in an 18-item questionnaire. The lowest, “1”, means that the person is totally dependent and the highest, “7”, means that the person is completely independent. These scores are used to measure the ability of the rehab patient in everyday tasks. These include physical tasks such as bathing, dressing, walking and eating. The scores also measure social and cognitive tasks such as problem solving, social interaction and comprehension of task.

Each score has an individual meaning. Table 1 shows what each score means and the classification given to a patient with that score (medfriendly.com).

Table 1- FIM Score Indications

Score	Indication	Classification
1	Total Assistance	Complete Dependence
2	Maximal Assistance	Complete Dependence
3	Moderate Assistance	Modified Dependence
4	Minimal Contact Assistance	Modified Dependence
5	Needs some supervision	Modified Dependence
	Or will ask for help	
6	Modified Independence	Independent
	independent but may use a	
	wheelchair or other assistive devices	
7	Total Independence	Independent

Trained professionals such as nurses or physical therapists assign FIM scores. Training to assign FIM scores typically involves educational videos, observations and a competency test. The scores are assigned at admission and discharge of rehabilitation. These scores are used to measure improvement from beginning of rehabilitation to the end of the program (www.medfriendly.com).

Many studies use these scores to measure the benefits of rehabilitation programs. The difference of the sum of the FIM scores at admission and the sum of the FIM scores at discharge can be analyzed directly, or used to calculate the change of FIM score per day or week (Fiedler *et al*,1996).

2.3 Medical Informatics

Research in healthcare relies on the collection of quality information. Trials and studies collect information in the hopes of using it to better treat patients or come up with improved techniques. Health Informatics is simply the "study of information and communication systems in healthcare" (Coiera, 2003). It deals with all aspects of health, including use and analysis of information. Bioinformatics is a field of medical informatics that deals with "information, data, and knowledge in the context of biological and biomedical research" (Chen *et al,* 2005). Text mining is also a subfield of medical informatics.

The term "medical informatics" has only been around for about thirty years. However, the concept has been around since healthcare began. When a doctor collects information to better treat their patients and to help treat future patients, that is considered health informatics.

The use of computers and new technology helps increase our knowledge of healthcare. Two growing fields of medical informatics are data mining and text mining. Data mining is generally used to explore and find patterns among structured data while text mining is used to extract information from unstructured text databases. These two fields overlap and use techniques from one another, as do many other fields in medical informatics.

2.4 Text Mining

Text mining extracts data from unstructured, nominal data. Nominal data refers to data that is categorical in nature without order. For example, the variable, gender, with categories male and female, would be considered nominal data. After previously unknown information is extracted, this text can be transformed into meaningful numbers so that statistical analysis can be conducted. Clusters on these data can then be formed to group similarities.

Text mining is currently used in the field of Bioinformatics to convert allele names into useful model data. Information about a certain allele is text mined to find the possible function of that allele. However, the allele name that is text mined is an alphanumeric code, such as “cyp2d6*4” allele. Even though the allele has numbers in the name, it can still be text mined since it is just a code. Therefore, text mining can be branched out to extract other useful information of health-related coded data, such as ICD-9 codes. These codes represent health conditions that a patient presents and are used for medical billing purposes.

Using a series of mathematical algorithms borrowed from data mining techniques, nominal data, such as ICD-9 codes, can be clustered and used in statistical models. These mathematical algorithms will be discussed in more detail in the Chapter 3.

2.5 ICD-9 Codes

International Classification of Diseases (ICD) codes are widely used in healthcare. This coding system started as a way for statisticians to keep track of how many deaths there were and what caused the deaths. The earliest classification system is credited to John Graunt on the London Bills of Mortality World Health Organization. He attempted to find the proportion of deaths from different diseases in children under the age of six. As time went on, other scientists and statisticians modified Graunt’s raw methods. The International Statistical Institute charged a committee in 1891 to prepare a classification of causes of death. Revisions to the classification are continuously made at periodic meetings. In 1946, the task of updating the system was passed on to the World Health Organization for international use. Currently, ICD-9 codes are the standard in medical coding. The version currently used is the ninth revision World Health Organization.

The ICD-9 is the “official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States” according to the Center for Disease Control. Each time a patient visits a physician, the diagnosis is given a numeric code. These codes are used for billing purposes and retained in medical records. All health conditions have corresponding numeric codes. These codes are also used for keeping track of mortality data, similar to the earlier classification systems. Table 2 gives examples of ICD-9 codes as well as their definitions.

Table 2: ICD-9 Codes

ICD-9 codes	Descriptions
244.9	Unspecified Hypothyroidism
250	Diabetes mellitus
276.8	Hypopotassemia (potassium deficiency)
278.01	Morbid obesity
285.9	Unspecified anemia
300	Neurotic disorders
305.1	Tobacco use disorder
357.2	Polyneuropathy in diabetes
401.9	Essential hypertension (unspecified)
428	Heart failure
496	Chronic airway obstruction
518.83	Chronic respiratory failure
530.81	Esophageal reflux
593.9	Unspecified disorder of the kidney or ureter
715.9	Osteoarthrosis, unspecified where
787.2	Dysphagia
799.3	Debility
v44.0	Tracheostomy
v45.01	Cardiac pacemaker

Along with ICD-9 codes, patients may also be given secondary codes, called comorbidity codes. Comorbidity is defined as “a pre-existing secondary diagnoses of the admitted patient” (Burchill, 2005). In our data, comorbidity codes are ICD-9 codes that represent secondary conditions. The comorbidity codes will be text mined and clustered to obtain a severity ranking.

CHAPTER 3

METHODOLOGY

3.1 Data

The data analyzed in this thesis was collected at Frazier Rehabilitation in Louisville, Kentucky. The data consist of 555 patients who were all receiving inpatient care at this facility for either cardiac or pulmonary disorders. There were 254(45.8%) patients with cardiac disorders and 301(54.2%) patients with pulmonary disorders. There were 247(44.5%) males and 308(55.5%) females ranging in age from 2 to 96 with a mean age of 70.93 (std dev=14.38). The patients were predominately Caucasian (476 out of 555 patients (85.8%)). There were also 78(14%) African-Americans and 1(0.2%) Pacific Islander enrolled in the study. The majority of patients were admitted from an acute care unit. The average length of stay at Frazier Rehabilitation was 13.29 days (std dev=6.81).

Each patient who was in this study had a main diagnosis code, which represents the primary reason for receiving care. The diagnosis codes allowed us to define a patient as cardiac or pulmonary in condition. No other conditions were included in the study. Each patient had a string of comorbidity codes, which were secondary to the diagnosis code. These strings range from 1 code to 10 codes, depending on the patient. These codes were used in the text mining stage of the analysis.

When the patients entered the rehabilitation center, they were assessed for functional independence and assigned FIM scores. They were measured in the activities mentioned in the background section. At discharge, they were reassessed to see if there was any improvement. These scores are listed in the dataset.

Overall, there are approximately 230 variables. The observations for some of the variables are not complete and several will not be used for this study, since they are concerned with financial elements. Some of the variables that were not used are methods of payment and types of insurance. Also, length of stay was not used in this study. Although length of stay and discharge destination are common response variables in rehabilitation studies, we chose to focus on functional independence measures to assess the benefits of cardiopulmonary rehabilitation. In the future, this data may be used to predict length of stay as a measure of rehabilitation. Since this study is dealing with human subjects, HIPAA Training and CITI Training were required. Also, the University IRB approved the study (approval letter attached in appendix A).

3.2 Text Mining

3.2.1 Background

The purpose of text mining is to extract information from observed similarities in textual data. These similarities can then be converted into meaningful numbers and used for data analysis. Common data mining techniques such as clustering and neural networks can be applied to these numerical indices for exploratory data analysis.

Text mining is used everyday in industries such as healthcare and marketing. For example, it is becoming very common to ask open-ended questions on surveys regarding a new product. This allows the customer to not be tied into a specific response, and to answer freely. Open-ended responses may bring up new ideas that the researcher was not aware of. After collecting the responses, text mining can then be conducted on the data to find similarities in the answers. Text mining helps to combine similar responses into meaningful clusters. These clusters can then be analyzed through statistical models to provide insight into how, for example, a company should market a new product.

There is, however, a misconception of what text mining is. There are two common definitions of text mining. One is previously discussed where we are trying to extract meaningful information and convert this information into numerical indices. Text mining can also refer to searching for specific text. This is widely used through internet search engines. A search will be conducted for a specific word or phrase and websites and documents containing that word or phrase will result. The Bioinformatics example provided earlier is an example of the second definition of text mining. This thesis will focus on the first definition of these two common types of text mining.

Data mining and text mining accomplish similar outcomes, but the input data define which process is used. Data mining is used when data are ordinal or interval. Data mining is concerned with identifying differences occurring within the data. For example, if “height” is the variable that is being mined, then we know that 6 feet 5 inches is larger than 5 feet 3 inches. However, text mining uses nominal data and makes no inferences about differences within the data. For example, if “pet” is the variable being text mined, then we cannot make a measurable difference between dog and cat. Even though dog and cat are two different animals, text mining does not assume that one is larger or better than the other.

3.2.2 Theory

The process of text mining involves a largely automated system of complex mathematical algorithms. To achieve the desired outcomes of text mining, a series of procedures must be conducted on the data. These procedures are: preprocessing, parsing, stemming, tagging, and transforming.

Preprocessing is the method of preparing the data for text mining. Very rarely does a dataset come in perfectly clean to the statistician. Preprocessing is a lengthy process to correct the data so that it can be imported into a computer program such as SAS and used for analysis. This can involve simply converting words into numeric codes or making sure that all of the observations have similar formats for the responses. For text mining, this may also mean taking individual words or phrases of each document and creating text strings.

Parsing is the selection of a variable that will be used to discover trends among the documents. This variable will be used in the text mining to find meaningful similarities among the observations. An example would be routing emails to a certain department. For instance, complaints may be sent to a government office and rerouted to the appropriate department. This is done by certain words and phrases being parsed in the email, which directs the message to the appropriate place.

Stemming refers to the process of taking similar terms and combining into one umbrella term. For example, if the word is “small”, then small, smaller and smallest would all fall under “small”. Also, synonyms can be input to combine terms, such as little for “small”. For this process, a list of synonyms usually needs to be specified. This can be done by the user and input to the text mining program. This is optional and is used to help reduce the number of words.

Tagging is utilized to exclude certain words, such as articles of speech. For instance, if the phrase was “milking the cow”, tagging can be used to exclude the word “the”. If the word has no importance to the phrase, then it can be excluded. Also, tagging can be used to determine what part of speech a word is. For example, if we had the phrases, “going to a show” and “let me show you something”, tagging would identify the meaning of show based on its usage in the phrase.

Last, transforming involves the actual creation of the matrices to convert the words into numerical indices. This includes creating a term by document frequency matrix, weighting the frequencies and transforming the frequencies.

A term by document frequency matrix is a matrix with terms by documents dimension. For example, Table 3 shows an example of a term-document matrix.

Table 3- Term-Document Matrix

Documents			
Term	D_1	D_2	D_n
T_1	$L_{1,1}$	$L_{1,2}$	$L_{1,n}$
T_2	$L_{2,1}$	$L_{2,2}$	$L_{2,n}$

where L $_{i,j}$ is the frequency weight for the ith term and jth document. This matrix gives the frequencies of the unstructured text. Instead of presenting the matrix as observations by variables, it is transposed to variables by observations, or in the case of text mining, terms by documents. Once this matrix is formed, then weights can be assigned to the frequencies. The three general options for weighting are:

$$\begin{aligned} &\text{none} \quad L_{ij} = a_{ij} \\ &\text{Binary} \quad L_{ij} = \begin{cases} 1 \text{ if term } i \text{ is in document } j \\ 0 \text{ otherwise} \end{cases} \\ &\text{Log} \quad L_{ij} = \log_2\left(a_{ij} + 1\right) \end{aligned}$$

where a_{ij} represents the frequency of term i in document j. The binary weighting system assigns a “1” if the term is in the document or a “0” if otherwise. The log weights even the distribution since certain words occur more often than other words. This basically means that the term with a higher frequency in the dataset receives a smaller weight and a term with a lower frequency receives a larger weight. These weights are referred to as local weights. However, the larger the document, the higher the frequencies will be for some words over smaller documents. Term weights are used to adjust the local weights to even the distribution of terms among the documents. This is demonstrated in the example later in this section.

Two types of term weights that are commonly used are entropy and normal. When entropy weighting is used, terms appearing exactly once in each document do not provide any new information about the data, and therefore give a weight of 0. If a term appears only one time in one document, then the weight is 1. The equation for entropy weight G_{Ei} is:

$$G_{Ei} = 1 + \frac{1}{\log_2(n)} \sum_j p_{ij} \log_2(p_{ij})$$

$$G_{Ei} = \begin{cases} 1 \text{ when } a_{ij} \text{ is 1 for 1 document} \\ 0 \text{ when } a_{ij} \text{ is always 1} \end{cases}$$

where *n* is the number of documents in the dataset and $p_{i,j} = \frac{a_{i,j}}{g_i}$ where $a_{i,j}$ is the frequency that term i appears in document j and g_i is the frequency that term i appears in the dataset. The second equation is just a special case of the first, specifically for when $a_{i,j}$ is always 1 or 1 for just one document.

Another commonly used type of term weight is Normal. This term weight is used when the frequency of terms within documents is more informative rather than the rarity of those terms. As illustrated in the equation below, as the frequency of term i across documents increases, the normal weight tends towards zero. For example, if a couple of documents have high frequencies of terms while the other documents do not have those terms, then the weight will be closer to zero than if the frequencies were similar across documents, and the terms will be viewed as unimportant, whereas they would be important in entropy. The equation for normal weight G_{Ni} is:

$$G_{Ni} = \frac{1}{\sqrt{\sum_j a_{i,j}^2}}$$

$$G_{Ni} = \begin{cases} \frac{1}{\sqrt{n}} \text{ when } a_{ij} \text{ is always 1} \\ 1 \text{ when } a_{ij} \text{ is 1 for 1 document} \end{cases}$$

The use of weights in the term by document frequency matrix helps convert the unstructured data into a structured matrix of numbers (SAS Text Miner Manual, 2004).

After weighting the term by document frequency matrix, there are numerous reasons it should be transformed. Often there are more terms than documents, so the matrix should be reduced in size. Also, since not all documents have the same terms, there are many empty cells. Lastly, we do not want to throw away documents or terms since they may contain useful information, but we would like to reduce the dimensionality by correcting for the empty cells. Therefore, we want to transform the matrix into something more manageable. A very common strategy for transforming the matrix is the Singular Value Decomposition (SVD) (SAS Text Miner Manual, 2004). This method is used for reducing dimensionality while still preserving information.

Singular Value Decomposition takes the term document frequency matrix, A, and expresses it in equivalent form as **A=UΣV,** where Σ is the diagonal matrix of singular values, σ, and U and V are orthogonal with right- and left-singular vectors. Left- and right-singular vectors means that for the singular value σ of A, AV= σ U and $A^TU=V$ where U is left-singular and V is right-singular. The transpose of the U matrix by A produces SVD document vectors and A by the transpose of the matrix V is the SVD term vectors. So, for an example with two documents and three terms, U^TA would look like:

$$\begin{array}{lcc} & \text{SVD1} & \text{SVD2} \\ \text{Document1} & U_{1,1}W_{1,1}+U_{2,1}W_{2,1}+U_{3,1}W_{3,1} & U_{1,2}W_{1,1}+U_{2,2}W_{2,1}+U_{3,2}W_{3,1} \\ \text{Document2} & U_{1,1}W_{1,2}+U_{2,1}W_{2,2}+U_{3,1}W_{3,2} & U_{1,2}W_{1,2}+U_{2,2}W_{2,2}+U_{3,2}W_{3,2} \end{array}$$

The matrix of AV^T would look like:

$$\begin{array}{lcc} & \text{SVD1} & \text{SVD2} \\ \text{Term1} & V_{1,1}W_{1,1} + V_{1,2}W_{1,2} & V_{2,1}W_{1,1} + V_{2,2}W_{1,2} \\ \text{Term2} & V_{1,1}W_{2v1} + V_{1,2}W_{2,2} & V_{2,1}W_{2,1} + V_{2,2}W_{2,2} \\ \text{Term3} & V_{1,1}W_{3,1} + V_{1,2}W_{3,2} & V_{2,1}W_{3,1} + V_{2,2}W_{3,2} \end{array}$$

where W_{ij} represents the matrix A. Therefore, with these matrices, SVD preserves the information from the original term by document frequency matrix.

To reduce the dimensionality, the singular values in the Σ matrix appear in descending order. A Heuristic SVD Cutoff Rule is also employed (SAS Text Miner Manual, 2004). This rule is based on the resolution and maximum number of values specified by the person performing text mining. To attempt the maximum dimension reduction of the SVD matrix that can be used without losing too much information, there are some guidelines to follow if using SAS Text Miner. First, if the document collection is larger, say in the thousands, the SVD dimension should be between 10 and 200. Otherwise, a dimension from 2 to 50 is more appropriate. If the goal is prediction, then a larger dimension, from 30-200, is appropriate. If the goal is clustering or exploration, then 2-50 is appropriate. If a high resolution is specified, then the Heuristic SVD cutoff rule uses the maximum dimension specified by the user. A higher resolution should summarize the data better, but requires more computing resources. If low or medium resolution is selected, then the Heuristic cutoff rule is applied (SAS Text Miner Manual, 2004). There must be a minimum of six documents, but at least 100 documents are recommended to get meaningful results. It is not clear as to whether these guidelines are general and would work with other text mining programs. The values generated by the Singular Value Decomposition can then be used for clustering the data.

To show an example, ten patients were randomly selected from the Frazier database. The variables in this small dataset were patient id and comorbidity text string. Using SAS Text Miner, the text strings were parsed. Using the SVD method, 5 dimensions was selected as the maximum and log weighting system was used. The following results are shown in Tables 4 and Table 5.

Table 4- Comorbidity Code strings

Combination	Patient number
V46.1 V44.0 V44.1 493.9 315.8	1
V45.82 728.3 250.5 362.01 250.6 536.3 337.1 593.9 V45.01	2
491.2 305.1	3
V45.01 414 250.4 583.81 285.9 278.01 593.9 401.9 728.3	4
V45.82 427.31 728.3 414 401.9 274.9 300 716.9 V43.65	5
403.91 V45.1 424.1 272.4 486 728.3	6
428 585 728.3 780.09 780.52 414 V45.01 715.9 530.81	7
728.3 V46.1 V44.0 564 787.2 V44.1	8
728.3 401.9 507 253.6 305.1 276.8 V15.81 780.79 305 V12.59	9
V45.81 728.3 285.1 401.9 272 564.1 366.9 722.93 288.8	10

Table 5- Comorbidity Codes and Entropy Weights

Term	Frequency	Weight
728.3	8	0.097
401.9	4	0.398
v45.01	3	0.523
414	3	0.523
v46.1	2	0.699
v44.0	2	0.699
v44.1	2	0.699
v45.82	2	0.699
593.9	2	0.699
305.1	2	0.699

The first table shows the text string associated with each patient. The second table shows the terms from the dataset. There were ten terms all together. ICD-9 code 728.3 appeared in eight of the text strings according to the frequency. Since it appeared the most, it has the lowest weight. The weights are assigned based on the frequency they appear in the dataset. The weights shown are the term weights, or entropy weights, obtained subsequent to the local log weights. The lower the frequency, the higher the term weight will be.

To dissect the example further, Table 6 shows the term by document matrix for one patient and Table 7 shows the terms and weights for that patient.

Table 6- Term-Document Matrix for Patient 4

Term-Document Matrix	
Term	Patient 4
V45.01	1
414	1
250.4	1
583.81	1
285.9	1
278.01	1
593.9	1
401.9	1
728.3	1

Table 7 – Comorbidity Codes and Weights for Patient 4

Term	Frequency	Weight
V45.01	3	0.523
414	3	0.523
250.4	1	1
583.81	1	1
285.9	1	1
278.01	1	1
593.9	2	0.699
401.9	4	0.398
728.3	8	0.097

Table 6 shows the term by document matrix for patient 4 in our example. As we can see, each term appears exactly once in patient 4’s text string, so the local weight is 1. Table 7 shows the terms and associated term weights. Since code 728.3 appears in patient 4’s text string and overall in 8 patients’ text strings, the term weight is 0.097. On the other hand, since code 278.01 only appears in patient 4’s text string and therefore only once in the data, it has a term weight of 1.

3.3 Clustering

The purpose of cluster analysis is to organize previously unclassified data into meaningful and useful information (SAS Text Miner Manual, 2004). Cluster analysis is an exploratory data analysis tool that aims to classify information without any explanation or interpretation (Statsoft.com). An example of clustering is shopping centers. A shopping center could be classified as a mall or strip mall, for example. A mall can further break down into types of stores, such as food, clothing, shoes, jewelry, etc. Clothing stores can then further be clustered as men's, women's or children's clothes. Therefore, the further "shopping center" is broken down, the more in common each store in that cluster has with each other. In text mining, there are two main types of clustering: Hierarchical and Expectation Maximization (EM). There are numerous other algorithms and techniques for clustering, but only hierarchical and EM will be covered since these are the two main options in SAS Text Miner.

Hierarchical clustering is an agglomerative process in which each cluster contains a single observation at first. This process works from the bottom-up, starting with each observation as its own cluster. Clustering proceeds iteratively by merging together the two least dissimilar (in distance) clusters together to form a new cluster, producing one less cluster at each iteration (Hastie *et al,* 2001). Iterations continue until only one cluster is left.

There are different algorithms for computing hierarchical clusters. These algorithms are based on the distances between clusters. Single Linkage clustering is based on the minimum distance between the elements of each cluster. Average Linkage clustering is based on the mean distance of the elements of each cluster. Lastly, Complete Linkage clustering uses the maximum distance between the elements of each cluster (Hastie *et al*,2001).

The Expectation Maximization clustering algorithm is very similar to the k-means algorithm. In general, these techniques are used for detecting clusters in the data and assign observations to these clusters. The EM algorithm makes an effort to determine the best distribution of k clusters given a mixture of k multivariate normal distributions of all the clusters. Unlike the k-means algorithm, expectation maximization can apply to categorical or continuous variables. In creating the best-fit probability distribution, the EM algorithm allows us to estimate the means and standard deviations of the clusters. A maximum number of clusters may be specified in which the EM algorithm will iterate until the optimal number of clusters is attained. Otherwise, a particular number of clusters can be specified in which the EM algorithm will iterate until the stopping criterion or specified number of clusters is reached.

Once the data is clustered, it may be necessary to introduce new patients or documents into the dataset. In this case, a scoring method may be useful. This can be done using statistical packages such as SAS Enterprise Miner. The scoring method computes the posterior probabilities given the data. For more information on practical implementation of scoring, see the SAS Text Miner Manual.

3.4 Text Mining in SAS Enterprise Miner

The data set with the text strings was imported into SAS 9.1. Then, using SAS Enterprise Miner, we first selected an input data set. Then, we connected SAS Text Miner to the data set. A step by step manual is attached in Appendix 2 to show how the Text mining was performed and demonstrate how to set the specifications that will be used in the analysis.

CHAPTER 4

ANALYSIS AND RESULTS

4.1 Text mining of comorbidity codes

To start, for each patient, text strings of the comorbidity codes were created. These strings were then used for text mining and clustering to create severity rankings. Using Text Miner, the variable selected to be parsed was the text strings. Since the strings contained numbers, the numbers option, as well as stemmed words as root form, was selected for identifying the terms. A singular value decomposition was selected for the transformation, with a low resolution and a maximum of 100 dimensions. A log weighting scale was used as well as entropy term weighting. Clustering was also performed using Text Miner. The clusters were based on the SVD dimensions using an expectation maximization approach. Initially, the number of clusters was arbitrarily selected to be seven. Table 8 shows the first five terms of each cluster as well as the frequency of the clusters.

Table 8 –Seven Clusters

Cluster #	Descriptive Terms	Frequency	Percentage
1	518.83, 733, 278.01, 250.6, 357.2	37	7%
2	799.3, 715.9, 530.81, 428, 250	122	22%
3	285.9, 276.8, 593.9, 787.2, 414	88	16%
4	357.2, 250.6, v45.81, v43.3, 496	94	17%
5	v68.61, 244.9, 250, 45.01, 401.9	113	20%
6	305.1, 787.91, 300, 799, 599	52	9%
7	v44.0, 44.1, 780.57, 787.2, 278.01	49	9%

As we can see, the frequencies are fairly evenly distributed among clusters. For comparison, the clustering procedure was repeated using the optimal number of clusters suggested by the EM algorithm, 12, and 10, 8, 6 and 4 clusters. The results are shown below in Table 9 -Table 12.

Table 9- 10 Clusters

10 Clusters			
Cluster #	Descriptive Terms	Frequency	Percentage
1	250.4, 250.6, 357.2, 278.01, 593.9	50	9%
2	733, 486, 244.9, 728.3, 428	70	13%
3	263.9, 8.45, 492.8, v44.0, 285.9	25	5%
4	v12.59, 482.41, 365.9, v45.81, 401.9	32	6%
5	311, 41.11, v09.0, v15.82, 486	44	8%
6	496, 427.31, v43.3, 250, 414	127	23%
7	799.3, 518.83, 427.89, 272.4, 428	39	7%
8	305.1, 300, 787.91, 799, 276.8	67	12%
9	v44.1, 599, 787.2, 530.81, 250	76	14%
10	416.8, 300, 780.57, 285.9, 428	25	5%

Table 10- 8 Clusters

8 Clusters			
Cluster #	Descriptive Terms	Frequency	Percentage
1	416.8, 300, 780.79, 285.9, 428	23	4%
2	278.01, v44.0, 780.57, 593.9, 428	65	12%
3	518.83, 272.4, 300, 305.1, 799	74	13%
4	v45.01, v12.59, 427.31, 414, 428	89	16%
5	250.6, v45.1, 357.2, v43.3, 403.91	65	12%
6	v45.89, 276.8, 285.9, 428, 427.31	85	15%
7	787.91, 244.9, 599, 401.9, 250	75	14%
8	v15.82, 787.2, 486, 496, v45.81	79	14%

Table 11- 6 Clusters

6 Clusters			
Cluster #	Descriptive Terms	Frequency	Percentage
1	v45.81, v43.3, 272.4, 250, 427.31	157	28%
2	300, 787.91, 311, 486, 401.9	89	16%
3	250.6, 250.4, 357.2, 278.01, 780.57	52	9%
4	425.4, v12.59, v45.01, 427.31, 414	72	13%
5	v44.1, v45.1, 787.2, v44.0, 403.91	48	9%
6	518.83, 733, 780.57, 530.81, 787.2	137	25%

Table 12- 4 Clusters

4 Clusters			
Cluster #	Descriptive Terms	Frequency	Percentage
1	357.2, 250.6, 780.57, 593.9, 428	132	24%
2	v44.0, 530.81, 787.2, 427.31, 428	172	31%
3	518.83, 733, 300, 799, 305.1	127	23%
4	244.9, 401.9, 250, 728.3, 486	124	22%

By comparing the different numbers of clusters chosen, four clusters appear to result in the most even distribution. Even though the codes were stemmed, SAS outputs a table listing all terms, not restricting to the stemmed term. From this, we can see that certain codes are clustered together. For example, in each varying number of clusters, codes 357 and 250 appear together. To help in the decision of how many clusters to use, a sensitivity analysis was conducted, and will be described in a later section.

4.2 Application

Using the Frazier Rehabilitation data, the text mining process was used. The preprocessing of the data included creating text strings of the comorbidity codes. There were between 1 and 10 comorbidity codes associated with each patient record. In Excel, these codes were combined into strings to prepare for text mining. This combination of comorbidity codes was the variable that was parsed. Since there were 555 patients, and each patient had a text string of length 1 to 10, we were dealing with a large document matrix.

As mentioned before, comorbidity codes are actually numbers. However, since these codes are nominal, text mining treats the codes as text since the codes represent words (Cerrito, 2004). These comorbidity codes were then stemmed. The first three digits of an ICD-9 code represent the diseases while the subsequent digits further classify the number (Cerrito *et al,* 2003). Therefore, all subsequent codes could be rolled into one term. For example, 410 represents acute myocardial infarction. Therefore, 410.11, 410.21 and so on would be combined with 410 since they stem from the same basic disease. The results were stemmed, but do not appear so in the output tables. Although the stemming process reduces the original 544 comorbidity codes to just 312, SAS Text Miner does not list stemmed terms in the cluster labels. Since we were not dealing with sentences, tagging was not used.

To transform the term-document matrix, a log weighting system was used for local weights and entropy weighting were used for term weights, although other choices may have been equally appropriate. We used entropy term weighting since we wanted to see the commonality of the codes among the strings. Also, singular value decomposition was used. A maximum dimension of 100 with low resolution was specified. Also, since the number of documents was not large by definition and the goal was clustering, by SAS Text Miner guidelines, a maximum dimension between 2 and 50 may have been more appropriate.

To cluster the data, the expectation maximization approach was used. Expectation maximization was used instead of hierarchical clustering because the procedure provides results that are more readily interpretable. Domain knowledge is required to provide meaning to the clusters through labeling. The EM clusters are more easily labeled (Cerrito, 2006).

4.3 Models for FIMS

4.3.1 Preliminary analysis

In the data, it appeared that there were several outliers, and the handling of this data is useful to discuss. In particular, there were several cases where the age, FIM score sums or length of stay appeared to be irregular. For this study, the average age was 70.93 years (Std. Dev=14.38). However, the minimum age was almost 2 years old. In fact, there were 10 patients under the age of eighteen. Since the majority of the patients were elderly, and Medicare was involved with this study, it was not clear whether to include children or not. After looking through literature, there was nothing stating whether children should or should not be included in cardiopulmonary rehabilitation studies. Therefore, to maintain the possible maximum number of observations, patients under the age of eighteen were included. Also, there were outliers in regards to length of stay. The average length of stay for a cardiopulmonary rehabilitation program is supposed to be two weeks. For this study, the average was 13.29 days, fairly close to the standard. However, the minimum was one day and the maximum was 46 days. Looking at these patients, there was no noticeable trend, such as patients with longer length of stays coming in with the highest or lowest FIM sum score at admission, nor were they of a certain age group. The outliers for length of stay did not appear to be related to age or condition. Again, since there was nothing that stated that these points should be discarded, the patients remained in the study. In the future, it would be beneficial to consult with a physician to try to identify causes for these outliers.

4.3.2 General Linear Models

We are trying to determine whether patients are receiving benefits from cardiopulmonary rehabilitation. We are hypothesizing that the sum of FIM scores at discharge can be predicted by the sum of admission FIM scores and the age of the patient, the severity of comorbidities and the interaction between the sum of admission FIM scores and age. Common sense tells us that if a patient enters the hospital with low FIM score sums and leaves with high FIM score sums, then the rehabilitation helped. The model also allowed for an interaction between age and admission FIM score sum. Lastly, the severity of the patients' illnesses will most likely affect the FIM scores. A severity ranking was assigned to each of the clusters obtained through text mining, as described below.

We first consider the conditions contained in each of the seven clusters. Table 13 shows the top five conditions in each cluster. The top five conditions are the five codes with the largest frequencies within that cluster.

Table 13- Seven Clusters and Definitions of Codes

Cluster 7	Definition of Codes
v44.0	Tracheostomy
v44.1	Gastrostomy
780.57	other & unspecified sleep apnea
787.2	Dysphagia
278.01	Morbid obesity
Cluster 6	
305.1	tobacco use disorder
787.91	Diarrhea
300	neurotic disorders
799	ill-defined and unknown cause of morbidity and mortality (or asphyxia)
599	other disorders of urethra and urinary tract
Cluster 5	
v58.61	long term (current) use of anticoagulants
244.9	Unspecified Hypothyroidism
250	Diabetes mellitus
v45.01	cardiac pacemaker
401.9	essential hypertension (unspecified)
Cluster 4	
357.2	polyneuropathy in diabetes
250.6	Diabetes with neurological manifestations
v45.81	aortocoronary bypass status
v43.3	heart valve
496	chronic airway obstruction
Cluster 3	
285.9	unspecified anemia
276.8	Hypopotassemia (potassium deficiency)
593.9	unspecified disorder of the kidney or ureter
787.2	Dysphagia
414	other forms of chronic ischemic heart disease
Cluster 2	
799.3	Debility
715.9	osteoarthrosis, unspecified where
530.81	esophageal reflux
428	heart failure
250	Diabetes mellitus
Cluster 1	
518.83	chronic respiratory failure
733	other disorders of bone and cartilage
278.01	Morbid obesity
250.6	Diabetes with neurological manifestations
357.2	polyneuropathy in diabetes

Assigning severity rankings based on the code definitions in Table 11 was out of the domain knowledge of the author. Therefore, an expert having knowledge of medical illnesses and terms (a pharmacist through personal communication) was contacted to rank these clusters in order of severity. The pharmacist has years of experience in ranking conditions and was a good resource for this issue. Using only the table of the clusters and codes/definitions, the consulted pharmacist was asked to rank the clusters on the basis of severity. The clusters were ranked as shown in Table 14.

Table 14- Ranking of Clusters

Cluster #	Rank
7	1
1	2
2	3
3	4
4	5
5	6
6	7

The higher the rank, the more severe the associated comorbidities, so a cluster with Rank=1 is most severe and a cluster with Rank=7 is least severe. Eventually, the ranking will be used for insurance reasons. For this study however, the actual ranking was not used in the model. The variable “rank” is just severity assigned to the clusters, which was still used as a nominal class variable. Therefore, the order assumed by the severity ranking was not used.

Results for the general linear model described above are shown in Table 15.

Table 15- GLM with Rank

Source	DF	SS	MeanSquare	F value	Pr>F
Model	9	148948.0549	16549.7839	97.03	<.0001
Error	545	92961.5378	170.5716		
Corrected Total	554	241909.5928			
R-Square	0.615718				

Source	DF	Type III SS	MeanSquare	F value	Pr>F
ADFIMSUMS	1	26472.13645	26472.13645	155.2	<.0001
AGE	1	1342.0686	1342.0686	7.87	0.0052
RANK	6	2113.80454	352.300076	2.07	0.0556
ADFIMSUMS*AGE	1	1570.95921	1570.95921	9.21	0.0025

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	13.2072707	6.8887224	1.92	0.0557
ADFIMSUMS	1.15348799	0.0925917	12.46	<.0001
AGE	0.27775312	0.0990205	2.81	0.0052
Rank 1	-8.41073782	2.6849614	-3.13	0.0018
Rank 2	-4.88775966	2.8107598	-1.74	0.0826
Rank 3	-1.88186127	2.1677982	-0.87	0.3857
Rank 4	-3.630964	2.2879452	-1.59	0.1131
Rank 5	-2.80426309	2.2657799	-1.24	0.2164
Rank 6	-3.79724762	2.1925592	-1.73	0.0839
Rank 7	0	.	.	.
ADFIMSUMS*AGE	-0.00421678	0.0013895	-3.03	0.0025

This model does not fully demonstrate that patients are benefiting from cardiopulmonary rehabilitation. However, this model does suggest that age has an impact on the sum of the FIM scores at discharge after adjusting for the sum of the FIM scores at admission. We can see from the parameter estimates that as FIM sum scores at admission and age increase, the discharge sum of FIM scores also increases. However, if both age and admission FIM sum score are higher, there will be a relatively lower increase. Also, the interaction term implies that the relationship between age and the discharge FIM scores depends on the admission FIM scores. Therefore, age does influence the benefits of the rehabilitation program, but is dependent on the FIM score sum with which the patients enter the program.

From this model, we can predict the discharge FIM scores sum. This could aid physicians who are deciding whether a person should enter rehabilitation. By assessing the patient's age, severity, and preliminary FIM scores, the physician could use this model to predict the possible benefits of the rehabilitation. The formula to predict for a particular cluster is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_x + \hat{\beta}_2 x_2 + \hat{\beta}_3 + \hat{\beta}_4 (x_1 * x_2)$ where $\hat{Y}$ is the prediction of FIM score sum at discharge, x_1 is the FIM sum score at admission and x_2 is age. $\hat{\beta}_0$ is the estimate of the intercept, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated coefficients associated with x_1 and x_2, respectively, $\hat{\beta}_3$ is the parameter estimate associated with severity ranking (different in each cluster), and $\hat{\beta}_4$ is the parameter estimate associated with the interaction term. Table 16 shows examples of predicting a patient's benefits.

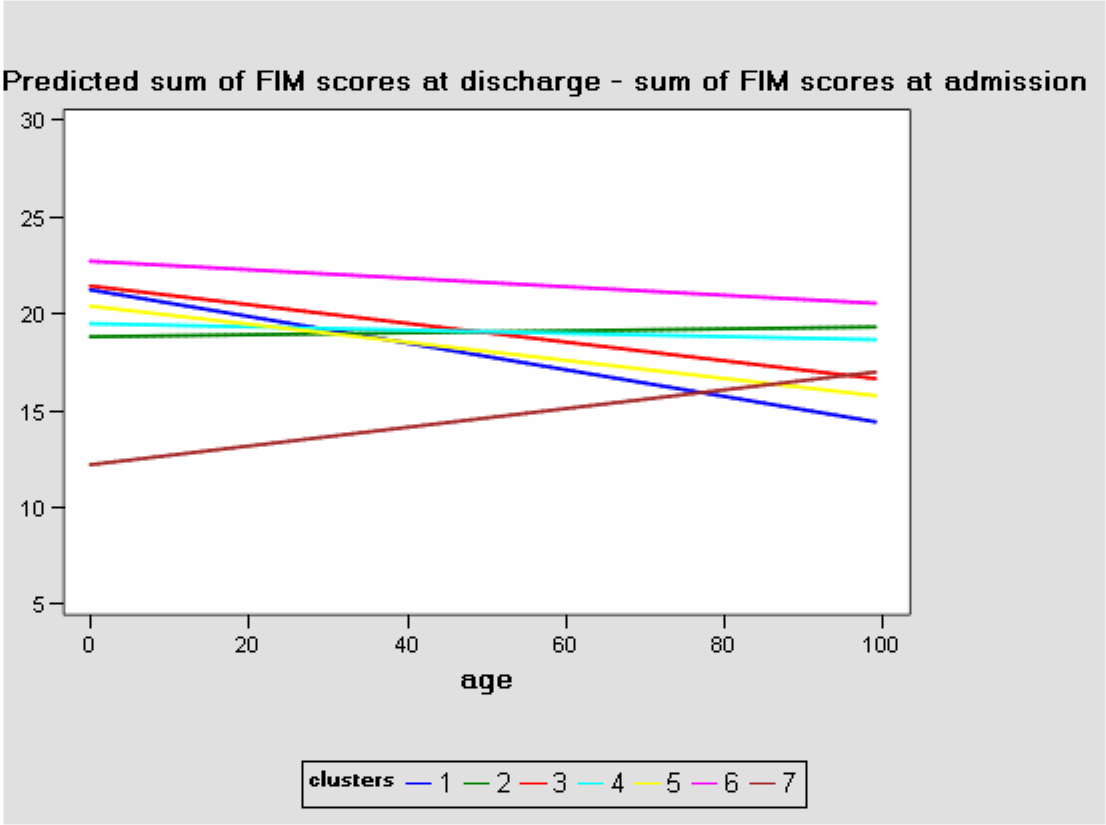
Table 16- predicting a Patient's Discharge FIM Score Sum

Patient	Age	Rank	Sums of FIM scores at admission	Predicted Discharge FIM score sum
1	69	1	54	70.53812898
2	58	2	76	93.50671497
3	37	3	47	68.48323082
4	91	4	41	66.41204364
5	75	5	54	76.44488584
6	20	6	40	57.73118152
7	46	7	60	83.5548821

From this table, we can see that each patient, regardless of age and rank, is predicted to leave the program with a higher FIM score sum. For these randomly selected patients, the predicted discharge FIM sum score is on average 20 points better than the admission FIM sum score.

We can also look at a graph of the predicted discharge FIM scores minus the sum of FIM scores at admission versus the age to see if there is a trend among the clusters (Figure 1).

Figure 1- Predicted Sums of FIM scores at discharge- FIM sum score at admission versus age



From the figure, we can see that older people in clusters 1, 3, 5 and 6 are predicted to have smaller improvement in sum of FIM scores from admission to discharge compared to younger ages. However, older people in clusters 2, 4 and 7 are predicted to have better improvement compared to the younger patients.

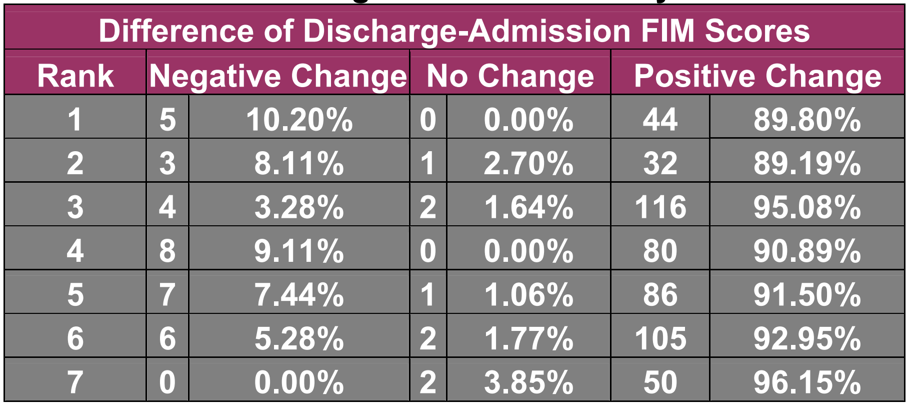
From the predictive model, we should be able to predict a new patient's sum of FIM scores at discharge based on their age, sum of FIM scores at admission and comorbidities. If this study was to reopen, we could score our data so that it would be easier to assign a patient to a certain cluster. This will eliminate the need for a doctor to classify the patient into a cluster. The patient should be assessed for FIM scores at admission. Then using the cluster they were assigned to, their age and the sum of the FIM scores, we should be able to predict their discharge FIM score.

The following section, 4.3.3, will use different methods to showcase that cardiopulmonary rehabilitation is potentially beneficial.

4.3.3 Other Methods for Demonstrating the Benefits of Rehabilitation

Another way to look at the data is to consider differences in FIM Scores (discharge minus admission). Table 15 shows the breakdown of frequencies in the classes “negative change”, “no change”, and “positive change” among the functional independence measure survey questions overall, and separately by severity rank clusters.

Table 17- Overall Changes in FIM Scores By Rank



Difference of Discharge-Admission FIM Scores						
Rank	Negative Change		No Change		Positive Change	
1	5	10.20%	0	0.00%	44	89.80%
2	3	8.11%	1	2.70%	32	89.19%
3	4	3.28%	2	1.64%	116	95.08%
4	8	9.11%	0	0.00%	80	90.89%
5	7	7.44%	1	1.06%	86	91.50%
6	6	5.28%	2	1.77%	105	92.95%
7	0	0.00%	2	3.85%	50	96.15%

Even though it appears there may be a trend, a Cochran-Armitage test for trend shows that positive change does not increase linearly with rank (One Sided Pr>Z = 0.2016).

Table 18- Rank 1

Difference of Discharge-Admission FIM Scores Rank 1				
Variable Differences in	Negative	No	Positive Change	% of Positive Change
	Change	Change		of out Total Change
Eating	6	24	19	38.76
Grooming	6	19	24	48.98
Dressing Upper Body	2	16	31	63.27
Dressing Lower Body	3	13	33	67.35
Controlling Bladder	8	20	21	42.85
Controlling Bowels	5	26	18	36.73
Transporting from Bed	4	17	28	57.14
Transporting to Toilet	3	17	29	59.18
Using Toilet	5	16	28	57.14
Walking or Using Wheelchair	2	9	38	77.54
Using Stairs	0	5	44	89.79
Transporting to Bathtub	1	7	41	83.66
Bathing	5	19	25	51.01
Expression	`	36	8	16.32
Social Interaction	2	35	12	24.49
Memory	4	29	16	32.65
Problem Solving	2	27	20	40.81
Comprehension	7	31	11	22.45

Table 19- Rank 2

Difference of Discharge-Admission FIM Scores Rank 2				
Variable Differences in	Negative	No	Positive Change	% of Positive Change
	Change	Change		of out Total Change
Eating	2	16	19	51.35
Grooming	1	15	21	56.76
Dressing Upper Body	1	12	24	64.87
Dressing Lower Body	2	7	28	75.67
Controlling Bladder	6	16	15	40.55
Controlling Bowels	5	18	14	37.83
Transporting from Bed	2	9	26	70.28
Transporting to Toilet	4	1	32	86.49
Using Toilet	4	7	26	70.26
Walking or Using Wheelchair	2	3	32	86.48
Using Stairs	2	4	31	8379.00
Transporting to Bathtub	1	3	33	89.19
Bathing	2	3	32	86.48
Expression	2	31	4	10.81
Social Interaction	1	31	5	15.51
Memory	5	23	9	24.33
Problem Solving	4	23	10	27.03
Comprehension	3	22	12	32.43

Table 20- Rank 3

Difference of Discharge-Admission FIM Scores Rank 3				
Variable Differences in	Negative Change	No Change	Positive Change	% of Positive Change of out Total Change
Eating	9	49	64	52.46
Grooming	3	31	88	72.13
Dressing Upper Body	3	37	82	66.22
Dressing Lower Body	4	18	100	81.97
Controlling Bladder	12	53	57	46.72
Controlling Bowels	13	49	60	49.18
Transporting from Bed	3	26	93	76.23
Transporting to Toilet	3	29	90	73.77
Using Toilet	5	21	96	78.68
Walking or Using Wheelchair	2	18	102	83.61
Using Stairs	4	18	100	81.97
Transporting to Bathtub	3	12	107	87.70
Bathing	7	20	95	77.87
Expression	9	83	30	24.59
Social Interaction	6	100	16	13.12
Memory	9	79	34	27.87
Problem Solving	7	75	40	32.79
Comprehension	7	87	28	22.95

Table 21- Rank 4

Difference of Discharge-Admission FIM Scores Rank 4				
Variable Differences in	Negative Change	No Change	Positive Change	% of Positive Change of out Total Change
Eating	4	36	48	54.54
Grooming	7	22	59	67.05
Dressing Upper Body	7	23	58	65.91
Dressing Lower Body	5	16	67	76.14
Controlling Bladder	10	38	40	45.46
Controlling Bowels	15	33	40	45.45
Transporting from Bed	5	20	63	71.59
Transporting to Toilet	6	19	63	71.60
Using Toilet	8	14	66	74.99
Walking or Using Wheelchair	0	15	73	82.96
Using Stairs	1	10	77	87.51
Transporting to Bathtub	2	10	76	86.36
Bathing	4	17	67	76.14
Expression	10	50	28	31.81
Social Interaction	4	72	12	13.63
Memory	7	63	18	20.46
Problem Solving	7	61	20	22.73
Comprehension	9	49	30	34.09

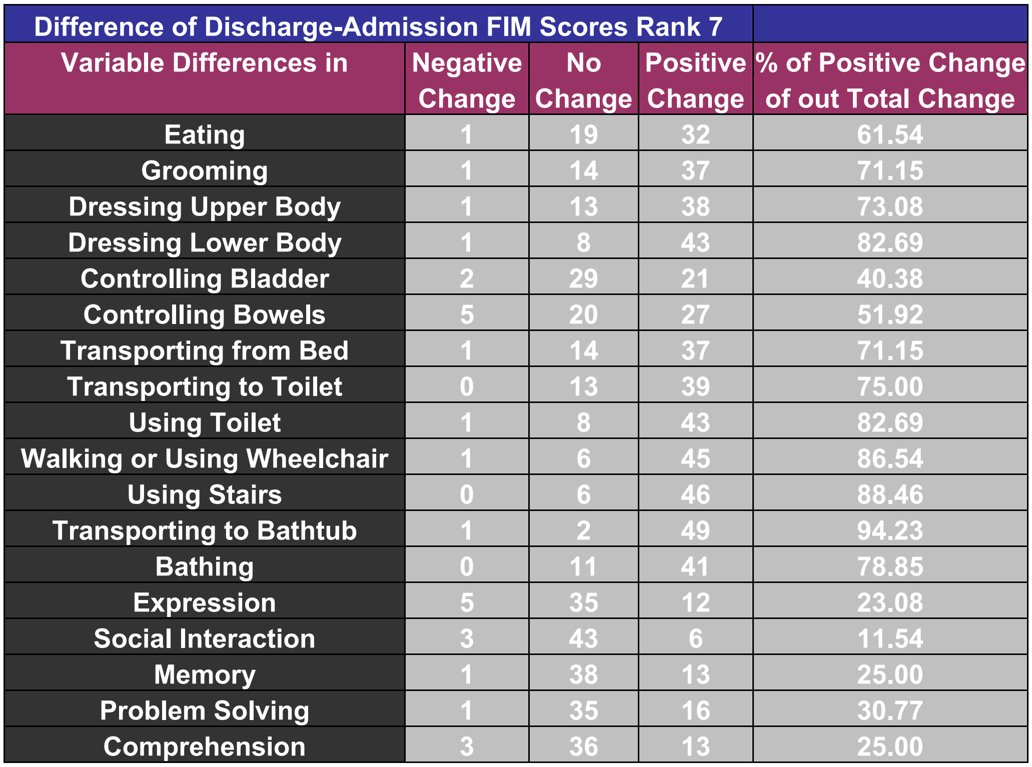
Table 22- Rank 5

Difference of Discharge-Admission FIM Scores Rank 5				
Variable Differences in	Negative	No	Positive Change	% of Positive Change
	Change	Change		of out Total Change
Eating	4	46	44	3.19
Grooming	9	23	62	65.95
Dressing Upper Body	4	24	66	70.21
Dressing Lower Body	4	13	77	81.91
Controlling Bladder	8	36	50	53.19
Controlling Bowels	10	36	48	51.06
Transporting from Bed	4	14	76	80.85
Transporting to Toilet	7	20	67	71.28
Using Toilet	7	14	73	77.66
Walking or Using Wheelchair	3	10	81	86.17
Using Stairs	4	7	83	88.30
Transporting to Bathtub	4	8	82	87.23
Bathing	5	20	69	73.40
Expression	6	63	25	26.60
Social Interaction	7	72	15	15.96
Memory	13	62	19	20.21
Problem Solving	11	59	24	25.53
Comprehension	9	60	25	26.60

Table 23- Rank 6

Difference of Discharge-Admission FIM Scores Rank 6				
Variable Differences in	Negative	No	Positive Change	% of Positive Change
	Change	Change		of out Total Change
Eating	8	55	50	44.25
Grooming	6	34	73	64.60
Dressing Upper Body	4	25	84	74.34
Dressing Lower Body	3	18	92	81.42
Controlling Bladder	9	58	46	40.71
Controlling Bowels	10	51	52	46.02
Transporting from Bed	8	21	84	74.34
Transporting to Toilet	5	20	88	77.88
Using Toilet	7	26	80	70.80
Walking or Using Wheelchair	3	8	102	90.27
Using Stairs	4	11	98	86.73
Transporting to Bathtub	3	15	95	84.07
Bathing	5	25	83	73.45
Expression	6	81	26	23.01
Social Interaction	5	89	19	16.81
Memory	13	75	25	22.12
Problem Solving	8	78	27	23.89
Comprehension	11	71	31	27.43

Table 24- Rank 7

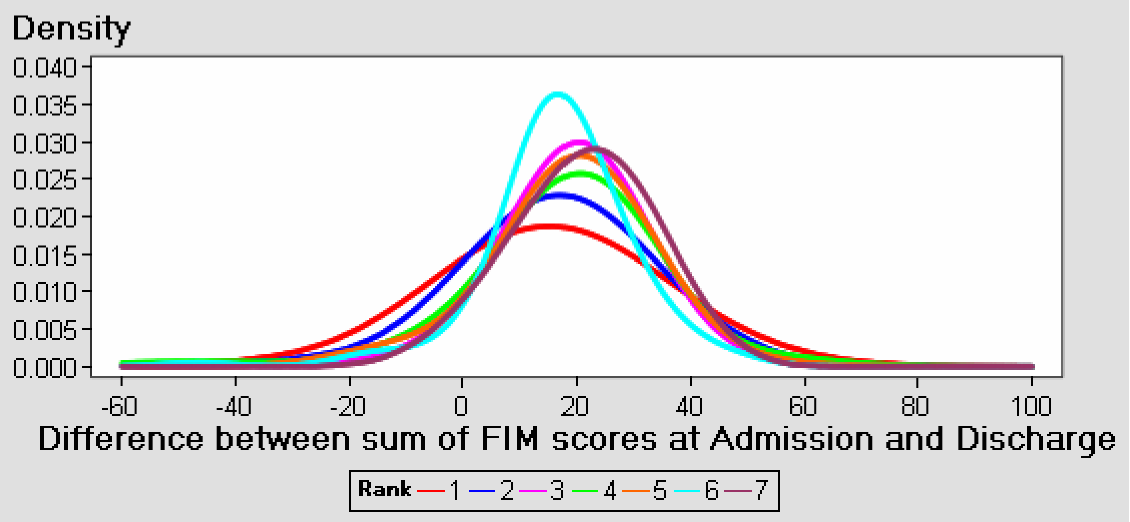


Difference of Discharge-Admission FIM Scores Rank 7				
Variable Differences in	Negative Change	No Change	Positive Change	% of Positive Change of out Total Change
Eating	1	19	32	61.54
Grooming	1	14	37	71.15
Dressing Upper Body	1	13	38	73.08
Dressing Lower Body	1	8	43	82.69
Controlling Bladder	2	29	21	40.38
Controlling Bowels	5	20	27	51.92
Transporting from Bed	1	14	37	71.15
Transporting to Toilet	0	13	39	75.00
Using Toilet	1	8	43	82.69
Walking or Using Wheelchair	1	6	45	86.54
Using Stairs	0	6	46	88.46
Transporting to Bathtub	1	2	49	94.23
Bathing	0	11	41	78.85
Expression	5	35	12	23.08
Social Interaction	3	43	6	11.54
Memory	1	38	13	25.00
Problem Solving	1	35	16	30.77
Comprehension	3	36	13	25.00

It is clear that almost all of the physical measures have a positive change due to rehabilitation regardless of severity of condition. The mental aspects appear to have no change. Even though the goal of the study was to look at improvement over physical measures, most cardiopulmonary programs include measurements of mental aspects (Lear et al, 2001). In keeping with what is found in literature about cardiopulmonary rehabilitation, the mental aspects were left in.

We can also look at the distribution of the ranks with differences between the sum of FIM scores at admission and discharge. To visualize the difference, kernel density estimation was conducted. Kernel density estimation approximates a hypothesized probability density function from the observed data. It is a nonparametric technique in which a known density function is averaged across the observed data points to create a smooth approximation. Using SAS code, densities of the difference between sum of FIM scores at admission and discharge were given for each patient. These densities were then separated according to rank. Figure 2 is the graphic results.

Figure 2- Kernel Density Estimation of Difference Between Sum of FIM Scores at Admission and Discharge



Looking at the peaks, or greatest density of each rank, there is a trend. As the difference in the sums increase, the density of the less severe ranks increase. However, it appears that rank 3, the pink line, could be switched with rank 6, the light blue line. It is perhaps not surprising that a perfect ordering of the severity rankings was not achieved, given that the rankings were assigned by a single expert. Therefore, it would have been better to consult a panel of experts with domain knowledge in medical conditions for ranking the clusters.

Also, by comparing the kernel density estimation of difference between sum of FIM scores at admission and discharge by rank, the difference is positive. This further indicates that cardiopulmonary rehabilitation is beneficial.

4.3.4 Sensitivity Analysis

Since the severity ranking will be based on the number of clusters chosen, we conducted a sensitivity analysis to assess the impact of choosing a different number of clusters on the results. Using SAS Enterprise Miner, the clustering process was repeated, specifying 4, 6, 8 and 10 clusters. Each resulting cluster variable was inserted in the general linear model described previously. Below are the results from general linear models using each cluster specification (4, 6, 8 or 10, as well as the original 7 clusters) to represent differences in comorbidities. A model without clusters was also considered.

Table 23- Ten Clusters General Linear Model

10 Clusters					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	148404.907	12367.0756	71.69	<.0001
Error	542	93504.6855	172.5179		
Corrected Total	554	241909.593			

R-Square	Coeff Var	Root MSE	DISFIMsum Mean
0.613473	13.66	13.13461	96.13694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sum of Fim Scores at Admission	1	24120.996	24120.996	139.82	<.0001
Age	1	1204.49845	1204.49845	6.98	0.0085
10 Cluster	9	1566.66061	174.0734	1.01	0.4314
Sum of Fim Scores at Admission*age	1	1276.3784	1276.3784	7.4	0.0067

Even though the model is significant at <.0001, having 10 clusters is not significant.

Table 26- Eight Clusters GLM

8 Clusters					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	148181.264	14818.1264	86	<.0001
Error	544	93728.329	172.2947		
Corrected Total	554	241909.593			

R-Square	Coeff Var	Root MSE	DISFIMsum Mean
0.612548	13.654	13.12611	96.13694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sum of Fim Scores at Admission	1	25986.6807	25986.6807	150.83	<.0001
Age	1	1431.94041	1431.94041	8.31	0.0041
8 clusters	7	1343.01713	191.85959	1.11	0.3528
Sum of Fim Scores at Admission*age	1	1449.49263	1449.49263	8.41	0.0039

Even though the model is significant at <.0001, having 8 clusters is not significant.

Table 27- 7 Clusters GLM

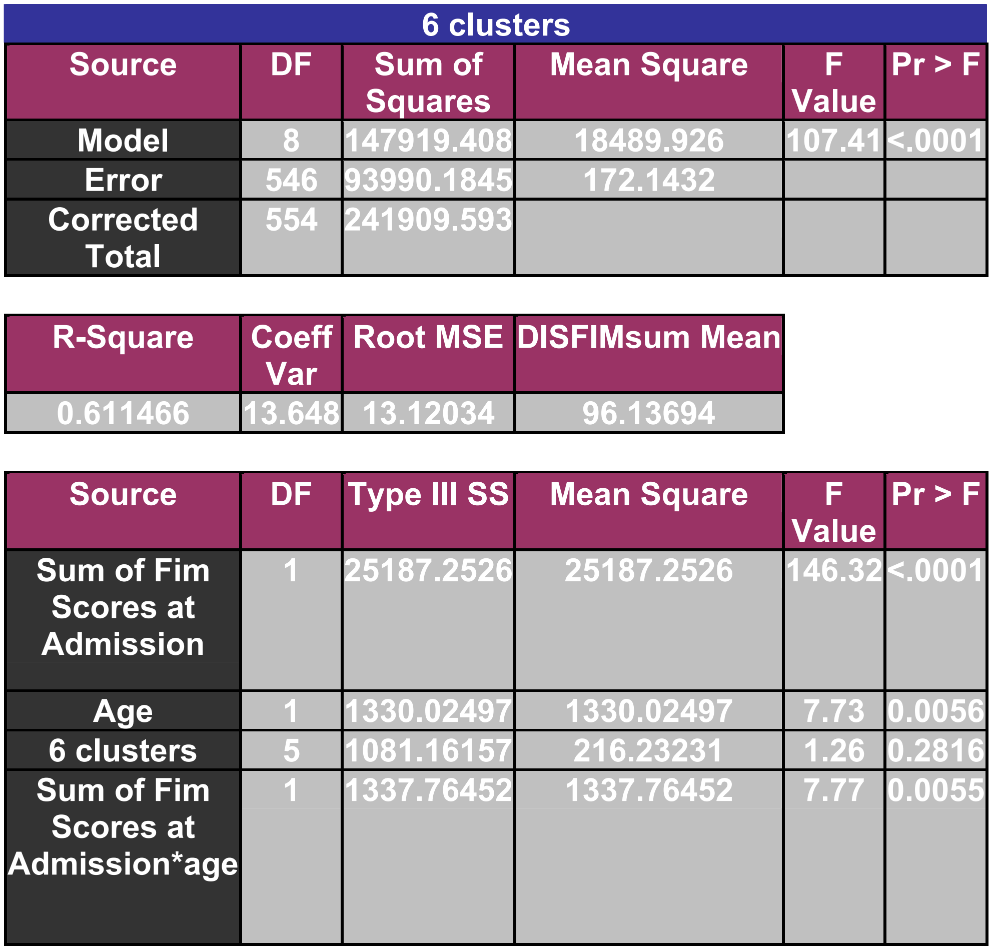
7 Clusters					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	148948.9	16549.88	97.03	<.0001
Error	545	92960.67	170.57		
Corrected Total	554	241909.6			

R-Square	Coeff Var	Root MSE	DISFIMsum Mean
0.615721	13.59	13.06025	96.13694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sum of Fim Scores at Admission	1	26488.86	26488.86	155.3	<.0001
Age	1	1349.147	1349.147	7.91	0.0051
7 clusters	6	2110.671	351.7786	2.06	0.056
Sum of Fim Scores at Admission*age	1	1571.329	1571.329	9.21	0.0025

The model here is significant at <.0001. Also, having seven clusters is marginally significant at α=0.1 level with a 0.056 p-value.

Table 28- 6 Clusters GLM



6 clusters

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	147919.408	18489.926	107.41	<.0001
Error	546	93990.1845	172.1432		
Corrected Total	554	241909.593			

R-Square	Coeff Var	Root MSE	DISFIMsum Mean
0.611466	13.648	13.12034	96.13694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sum of Fim Scores at Admission	1	25187.2526	25187.2526	146.32	<.0001
Age	1	1330.02497	1330.02497	7.73	0.0056
6 clusters	5	1081.16157	216.23231	1.26	0.2816
Sum of Fim Scores at Admission*age	1	1337.76452	1337.76452	7.77	0.0055

Even though the model is significant at <.0001, having 6 clusters is not significant.

Table 29- 4 Clusters GLM

4 Clusters					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	147243.196	24540.53	142.06	<.0001
Error	548	94666.3969	172.7489		
Corrected Total	554	241909.593			

R-Square	Coeff Var	Root MSE	DISFIMsum Mean
0.60867	13.672	13.1434	96.13694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sum of Fim Scores at Admission	1	25573.5905	25573.59	148.04	<.0001
Age	1	1159.39327	1159.393	6.71	0.0098
4 Clusters	3	404.94916	134.9831	0.78	0.5047
Sum of Fim Scores at Admission*age	1	1198.29505	1198.295	6.94	0.0087

Even though the model is significant at <.0001, having 4 clusters is not significant.

Table 30- GLM with no clusters

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	146838.25	48946.0822	283.7	<.0001
Error	551	95071.346	172.5433		
Corrected Total	554	241909.59			

R-Square	Coeff Var	Root MSE	DISFIMsum Mean
0.606996	13.663	13.13557	96.13694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sum of FIM scores at Admission	1	25786.639	25786.6393	149.5	<.0001
Age	1	1157.6604	1157.66039	6.71	0.01
Sum of FIM scores at Admission*age	1	1182.7734	1182.77338	6.85	0.009

After running a sensitivity analysis, we can see that the number of clusters does not have a large impact on the model. The Expectation Maximization classified twelve clusters to be optimal. However, with the exception of four clusters, 10, 8, 7 and 6 clusters change the p-values of the age and the interaction term marginally. The twelve clusters chosen by the EM procedure did not affect the model and produced similar outputs to the 10, 8, 7 and 6 clusters. This demonstrates that the predictors are not relying on the number of clusters chosen, which is positive. Also, the R-square value does not change much across the models, even if the clusters are excluded from the model. The significance of age and the interaction between age and sum of FIM scores at admission does not change substantially when the number of clusters is varied or when clusters are left out of the model altogether. Therefore, it seems arbitrary as to which number of clusters is chosen.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Positives and Negatives of Text Mining

Overall, text mining seems to be a very useful tool in analyzing textual or nominal data. However, this process cannot be done by hand. One of the drawbacks is that a sophisticated computer program must be used for text mining. Programs, such as SAS, offer text mining, but if SAS is not the program someone is used to, it can be difficult to learn. Also, packages that offer text mining tend to be expensive. Also, since this process is done by the computer, intermediate steps such as the singular value decomposition is all taken for granted. It must be assumed that the program was written correctly and that the singular value decompositions are accurate. Additionally, the time spent on deciding which clusters to use can be great. In SAS Text Miner, options such as k-nearest neighbor clustering are not available. It may be more beneficial to consider using the clustering node in Enterprise Miner to cluster the data.

However, the process taken to cluster the comorbidity codes can be done using other standard statistical software packages. To process the data in a similar fashion to text mining, there are four steps that need to take place. First, the ICD-9 codes should be truncated to remove the last two digits. This is similar to the stemming process, but can easily be done in many packages. Next, these codes should be transformed, both locally and across documents. This transformation will allow a dimension reduction process, such as singular value decomposition, to take place. Most packages have the capability to perform singular value decomposition. Last, the codes can then be clustered and used in statistical analysis.

Text mining has expanded the statistical tools for analyzing data. Instead of analyzing surveys by hand, health institutions can now have a computer program do this and save time for patients and other work. Also, if analyzing is done by hand, the number of surveys is usually kept small whereas text mining can analyze many surveys in a fraction of the time. Also, with the development of text mining, we are now able to apply data mining techniques, such as clustering and neural networks, to textual and nominal data.

In the past, logistic regression was the standard technique used to produce severity rankings based on ICD-9 codes Cerrito *et al* (2003). The way this was done was by combining chi-square tests and stepwise logistic regression to analyzing the ICD-9 codes. First, a binary term-document matrix was created by assigning a 0 if a patient does not have that ICD-9 code and 1 otherwise as dummy codes. Next, the chi-square was used to reduce the binary term-document matrix by eliminating all non-significant codes. Stepwise logistic regression was then used to further reduce the matrix. The use of logistic regression means that the input variables are assumed to be independent of each other. Therefore, two conditions cannot have any relation such as obesity and diabetes. Also, the model assumes that there is uniformity among the assigning of the ICD-9 codes, which rarely happens with multiple providers. The outcome variable in this model is usually mortality. Once the model is complete, each patient has a sum of weights where the remaining ICD-9 code dummy variables are equal to 1. These weights are used to assign a predictive value of mortality. In other words, if the patient's weights sum is greater than a cutoff point, then mortality is predicted. Therefore, the more codes a patient has that are assigned 1, then they will have a higher predicted risk of mortality. Furthermore, this suggests that the more codes a patient has in general, the more coded 1's he or she will have. In general, this model favors persons with more ICD-9 codes (Health Grades, 2006). If this approach was applied with discharge FIM score as the outcome, stepwise linear regression, rather than stepwise logistic regression, would be used.

The benefits of text mining over the use of stepwise linear regression are that: text mining does not need the assumption of uniformity among the data entry, it does not require linearity, and the codes can be related.

With the advancement of text mining, more and more quality of life issues may be examined.

5.2 Discussion

Overall, it has been shown that cardiopulmonary rehabilitation has been beneficial to cardiopulmonary patients. The general linear models demonstrated that it does not matter which number of clusters is chosen for the model. The clusters appear to have no influence on the other predictors, so seven clusters were arbitrarily chosen. This model, however, did not fully demonstrate that patients were benefiting from the program.

Since we cannot rely solely on the model, other methods were used to show the potential benefits of cardiopulmonary rehabilitation. By simply looking at frequency counts, it was shown that among all physical measurements, the differences in the FIM scores from admission to discharge were positive. This helps suggest the benefits of the cardiopulmonary rehabilitation. Furthermore, kernel density estimations of the differences in the FIM scores from admission to discharge by rank were overall positive. Also, the KDE demonstrated that there is not only a trend among the ranks, but that the differences in the sums are positive. These methods helped confirm that patients are potentially receiving benefits from cardiopulmonary rehabilitation.

5.3 Conclusion

Text mining and clustering are very effective for creating useful variables. In the health industry, clinical trials are sometimes kept smaller so that text analysis is not too time consuming. With SAS Enterprise Miner, text strings can be created and clustered for valuable statistical analysis. In this study, comorbidity codes were mined and clustered to assign patient rank severity. Formation of the clusters to reflect patient severity may be beneficial in the modeling process. With the help of text mining, we were able to show that by using the sum of admission FIM scores and age of patients can help predict the sum of the FIM scores at discharge. With the help of other statistical methods, we were able to show that patients are benefiting from cardiopulmonary rehabilitation.

REFERENCES

Burchill, Charles. “Complications and Comorbidities”. University of Manitoba. 2005. http://www.umanitoba.ca/centres/mchp/concept/dict/comorb_compl/comp_comorb.html

Center for Disease Control. www.cdc.gov

Cerrito, Patricia, Baden, Antonio & Cox, James. “The Application of Text Mining Software to Examine Coded Information”. http://www.siam.org/meetings/sdm03/proceedings/sdm03_25.pdf. 2003

Cerrito, Patricia. “Inside Text Mining”. *Health Management Technology.* March 2004.

Cerrito, Patricia. Introduction to SAS Enterprise Miner and Data Mining. SAS Press, Cary, NC. Projected Release October 2006.

Chen, Hsinchun, Fuller, Sherrilynne S., Friedman, Carol, Hersh, William. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Spring Science+Business Media, New York, NY. 2005. pp 69, 401-419.

Coiera, Enrico. Guide to Health Informatics. Hodder Arnold Publications. 2003.

Fiedler, Roger, et.al. “The Uniform Data System for Medical Rehabilitation: Report of First Admissions for 1994”. *American Journal of Physical Medicine and Rehabiliation*. Volume 75(2), March/April 1996, pp 125-129.

Foley, Margaret Wiley et al. Cardiopulmonary Rehabilitation: Basic Theory and Application (Contemporary Perspectives in Rehabilitation). F.A. Davis Company, Philadelphia. 1993. pp 1-10.

Functional Independence Measure. http://www.medfriendly.com/functionalindependencemeasure.html

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY. 2001. pp 472-478.

Healthgrades.com. “HealthGrades Fourth Annual Hospital Quality and Clinical Excellence Study”. *Healthgrades.com*. February 2006.

Lear, Scott, Ignaszewski, Andrew. “Cardiac rehabilitation: a comprehensive review”. Current Control Trials Cardiovascular Medicine. 2001; 2(5): 221–232.

Sansone, Giorgio et. al. “Analysis of FIM Instrument Scores for Patients Admitted Inpatient Cardiac Rehabilitation Program”. Archives of Physical Medicine and Rehabilitation. Vol 83, April 2002. pp 506-512.

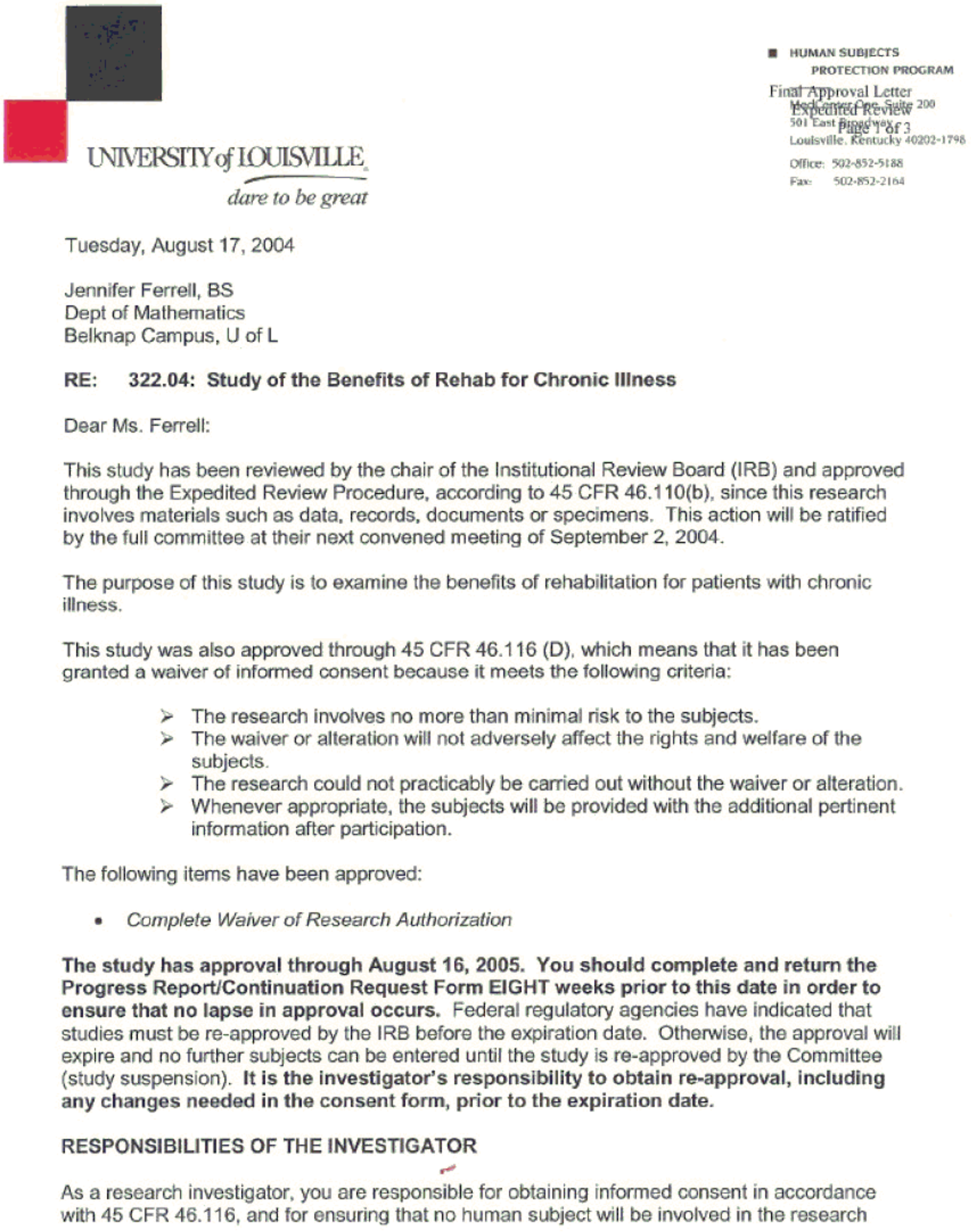
SAS Text Miner Manual. SAS Institute, Cary, NC. 2004

The Statistics Homepage. http://www.statsoft.com/textbook/stathome.html. Statsoft,Inc. 1984-2005.

Thearling, Kurt. “Scoring Your Customers”. http://www.thearling.com/text/scoring/scoring.htm

World Health Organization. “History of the development of the ICD”. http://www.who.int/classifications/icd/en/HistoryOfICD.pdf

APPENDIX 1



UNIVERSITY of LOUISVILLE
dare to be great

HUMAN SUBJECTS
PROTECTION PROGRAM
Final Approval Letter
Expedited Review
Page 1 of 3
MedCenter One, Suite 200
501 East Broadway
Louisville, Kentucky 40202-1798
Office: 502-852-5188
Fax: 502-852-2164

Tuesday, August 17, 2004

Jennifer Ferrell, BS
Dept of Mathematics
Belknap Campus, U of L

RE: 322.04: Study of the Benefits of Rehab for Chronic Illness

Dear Ms. Ferrell:

This study has been reviewed by the chair of the Institutional Review Board (IRB) and approved through the Expedited Review Procedure, according to 45 CFR 46.110(b), since this research involves materials such as data, records, documents or specimens. This action will be ratified by the full committee at their next convened meeting of September 2, 2004.

The purpose of this study is to examine the benefits of rehabilitation for patients with chronic illness.

This study was also approved through 45 CFR 46.116 (D), which means that it has been granted a waiver of informed consent because it meets the following criteria:

- ➢ The research involves no more than minimal risk to the subjects.
- ➢ The waiver or alteration will not adversely affect the rights and welfare of the subjects.
- ➢ The research could not practicably be carried out without the waiver or alteration.
- ➢ Whenever appropriate, the subjects will be provided with the additional pertinent information after participation.

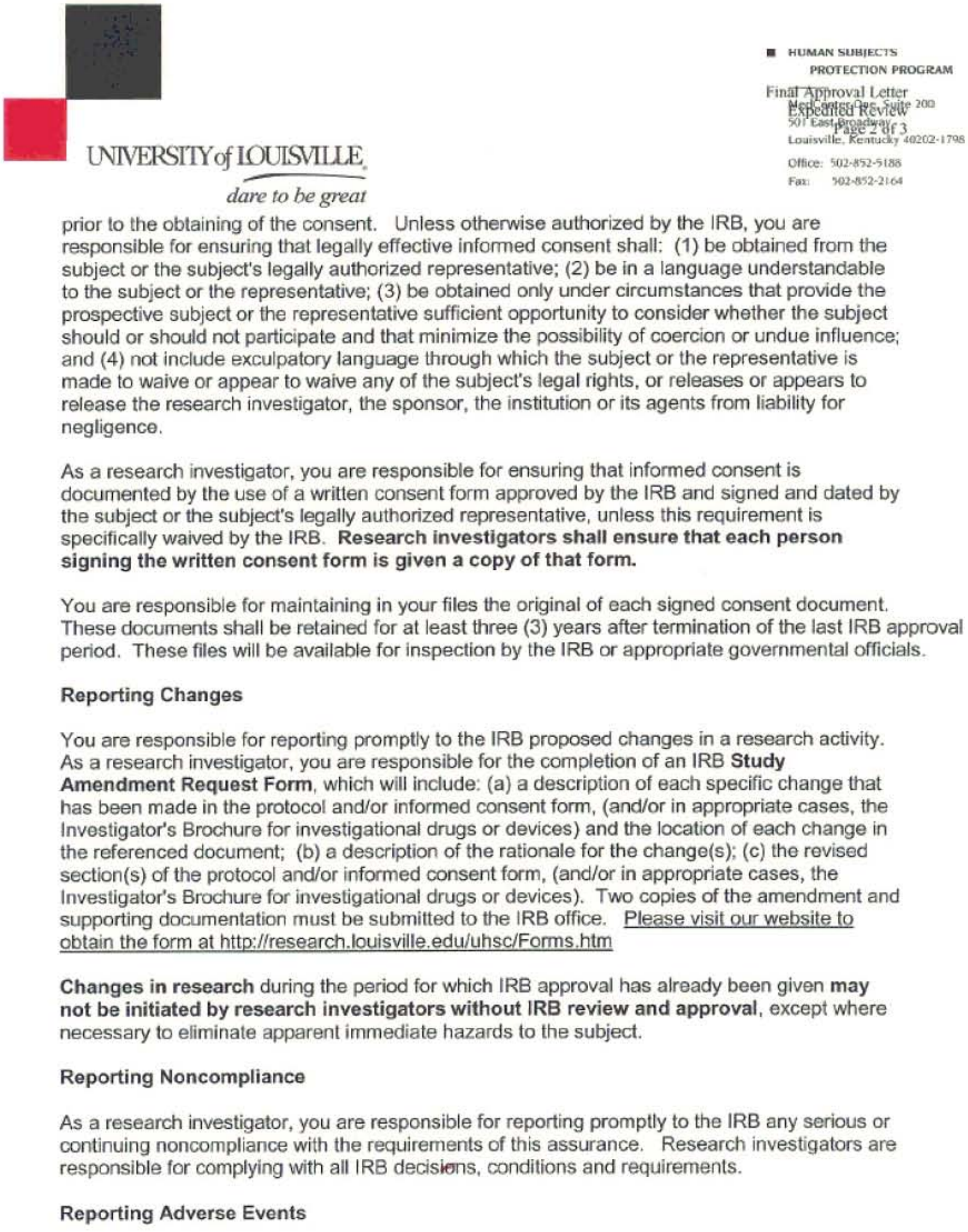
The following items have been approved:

- *Complete Waiver of Research Authorization*

The study has approval through August 16, 2005. You should complete and return the Progress Report/Continuation Request Form EIGHT weeks prior to this date in order to ensure that no lapse in approval occurs. Federal regulatory agencies have indicated that studies must be re-approved by the IRB before the expiration date. Otherwise, the approval will expire and no further subjects can be entered until the study is re-approved by the Committee (study suspension). **It is the investigator's responsibility to obtain re-approval, including any changes needed in the consent form, prior to the expiration date.**

RESPONSIBILITIES OF THE INVESTIGATOR

As a research investigator, you are responsible for obtaining informed consent in accordance with 45 CFR 46.116, and for ensuring that no human subject will be involved in the research



UNIVERSITY of LOUISVILLE
dare to be great

HUMAN SUBJECTS
PROTECTION PROGRAM
Final Approval Letter
Expedited Review
Page 2 of 3
MedCenter One, Suite 200
501 East Broadway
Louisville, Kentucky 40202-1798
Office: 502-852-5188
Fax: 502-852-2164

prior to the obtaining of the consent. Unless otherwise authorized by the IRB, you are responsible for ensuring that legally effective informed consent shall: (1) be obtained from the subject or the subject's legally authorized representative; (2) be in a language understandable to the subject or the representative; (3) be obtained only under circumstances that provide the prospective subject or the representative sufficient opportunity to consider whether the subject should or should not participate and that minimize the possibility of coercion or undue influence; and (4) not include exculpatory language through which the subject or the representative is made to waive or appear to waive any of the subject's legal rights, or releases or appears to release the research investigator, the sponsor, the institution or its agents from liability for negligence.

As a research investigator, you are responsible for ensuring that informed consent is documented by the use of a written consent form approved by the IRB and signed and dated by the subject or the subject's legally authorized representative, unless this requirement is specifically waived by the IRB. **Research investigators shall ensure that each person signing the written consent form is given a copy of that form.**

You are responsible for maintaining in your files the original of each signed consent document. These documents shall be retained for at least three (3) years after termination of the last IRB approval period. These files will be available for inspection by the IRB or appropriate governmental officials.

Reporting Changes

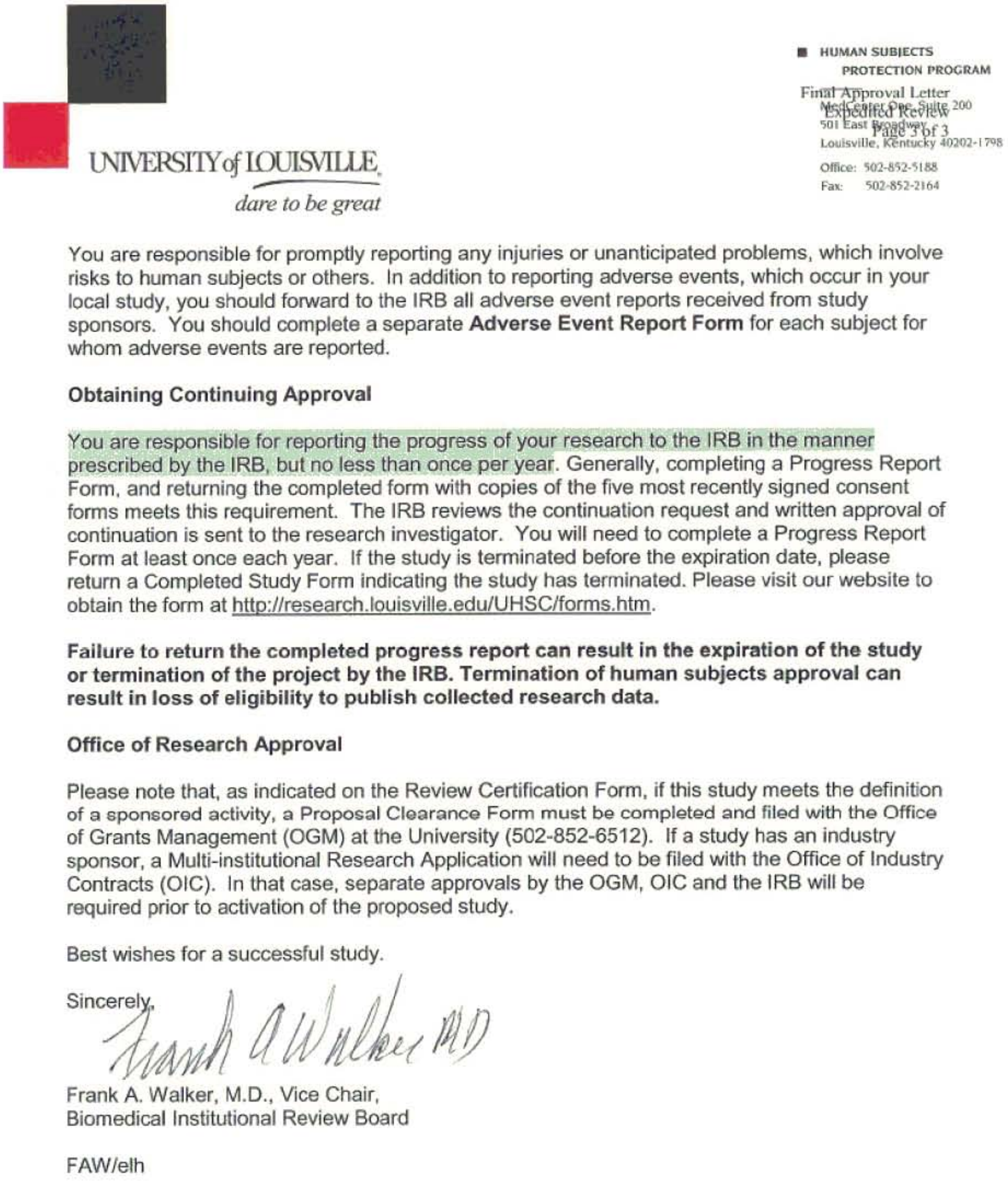
You are responsible for reporting promptly to the IRB proposed changes in a research activity. As a research investigator, you are responsible for the completion of an IRB **Study Amendment Request Form**, which will include: (a) a description of each specific change that has been made in the protocol and/or informed consent form, (and/or in appropriate cases, the Investigator's Brochure for investigational drugs or devices) and the location of each change in the referenced document; (b) a description of the rationale for the change(s); (c) the revised section(s) of the protocol and/or informed consent form, (and/or in appropriate cases, the Investigator's Brochure for investigational drugs or devices). Two copies of the amendment and supporting documentation must be submitted to the IRB office. Please visit our website to obtain the form at http://research.louisville.edu/uhsc/Forms.htm

Changes in research during the period for which IRB approval has already been given **may not be initiated by research investigators without IRB review and approval**, except where necessary to eliminate apparent immediate hazards to the subject.

Reporting Noncompliance

As a research investigator, you are responsible for reporting promptly to the IRB any serious or continuing noncompliance with the requirements of this assurance. Research investigators are responsible for complying with all IRB decisions, conditions and requirements.

Reporting Adverse Events



UNIVERSITY of LOUISVILLE

dare to be great

HUMAN SUBJECTS
PROTECTION PROGRAM

Final Approval Letter
Expedited Review
Page 3 of 3

MedCenter One, Suite 200
501 East Broadway
Louisville, Kentucky 40202-1798

Office: 502-852-5188
Fax: 502-852-2164

You are responsible for promptly reporting any injuries or unanticipated problems, which involve risks to human subjects or others. In addition to reporting adverse events, which occur in your local study, you should forward to the IRB all adverse event reports received from study sponsors. You should complete a separate **Adverse Event Report Form** for each subject for whom adverse events are reported.

Obtaining Continuing Approval

You are responsible for reporting the progress of your research to the IRB in the manner prescribed by the IRB, but no less than once per year. Generally, completing a Progress Report Form, and returning the completed form with copies of the five most recently signed consent forms meets this requirement. The IRB reviews the continuation request and written approval of continuation is sent to the research investigator. You will need to complete a Progress Report Form at least once each year. If the study is terminated before the expiration date, please return a Completed Study Form indicating the study has terminated. Please visit our website to obtain the form at http://research.louisville.edu/UHSC/forms.htm.

Failure to return the completed progress report can result in the expiration of the study or termination of the project by the IRB. Termination of human subjects approval can result in loss of eligibility to publish collected research data.

Office of Research Approval

Please note that, as indicated on the Review Certification Form, if this study meets the definition of a sponsored activity, a Proposal Clearance Form must be completed and filed with the Office of Grants Management (OGM) at the University (502-852-6512). If a study has an industry sponsor, a Multi-institutional Research Application will need to be filed with the Office of Industry Contracts (OIC). In that case, separate approvals by the OGM, OIC and the IRB will be required prior to activation of the proposed study.

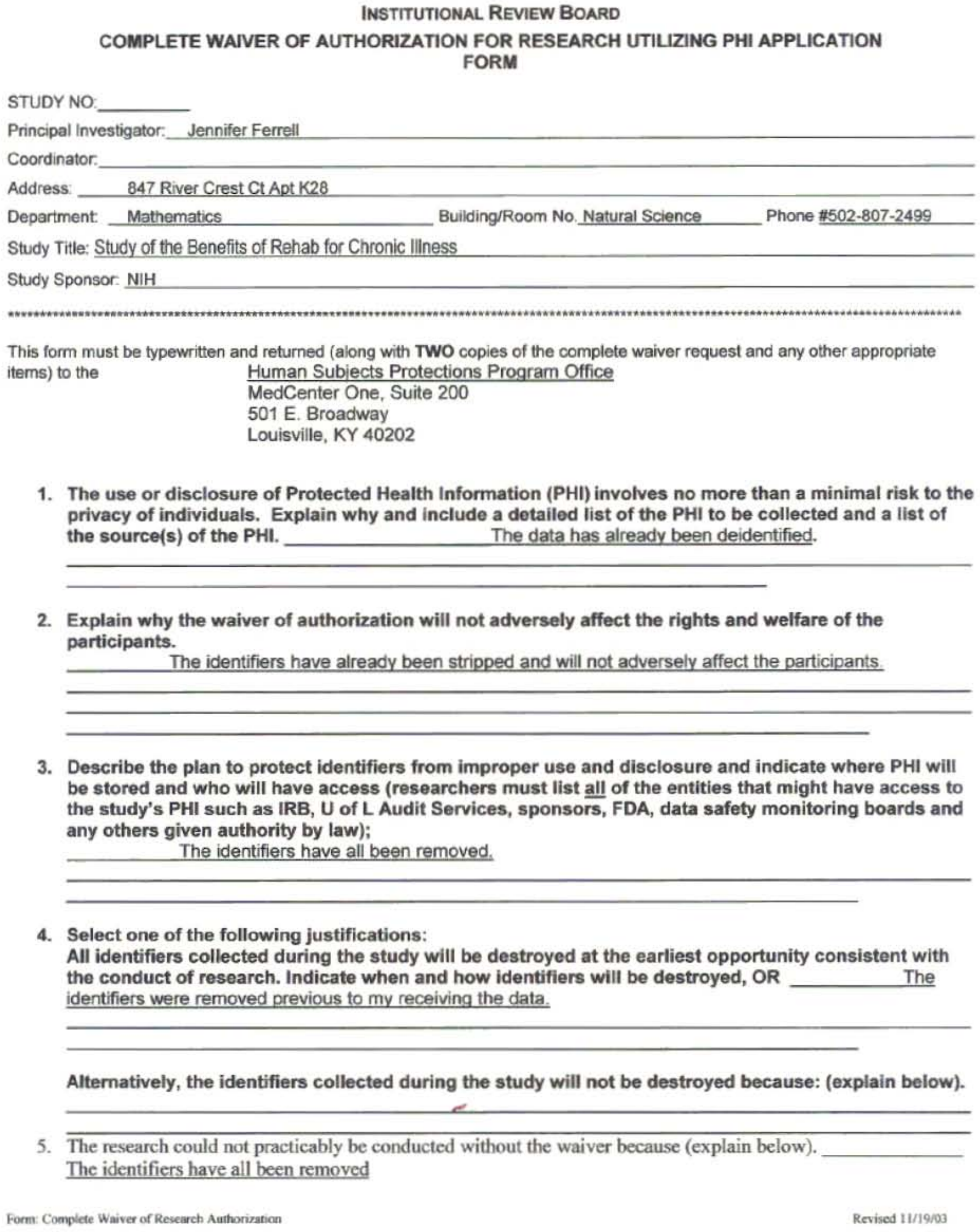
Best wishes for a successful study.

Sincerely,

Frank A Walker MD

Frank A. Walker, M.D., Vice Chair,
Biomedical Institutional Review Board

FAW/elh



INSTITUTIONAL REVIEW BOARD

COMPLETE WAIVER OF AUTHORIZATION FOR RESEARCH UTILIZING PHI APPLICATION FORM

STUDY NO:_________

Principal Investigator: Jennifer Ferrell

Coordinator:

Address: 847 River Crest Ct Apt K28

Department: Mathematics Building/Room No. Natural Science Phone #502-807-2499

Study Title: Study of the Benefits of Rehab for Chronic Illness

Study Sponsor: NIH

****

This form must be typewritten and returned (along with **TWO** copies of the complete waiver request and any other appropriate items) to the

Human Subjects Protections Program Office
MedCenter One, Suite 200
501 E. Broadway
Louisville, KY 40202

1. **The use or disclosure of Protected Health Information (PHI) involves no more than a minimal risk to the privacy of individuals. Explain why and include a detailed list of the PHI to be collected and a list of the source(s) of the PHI.** The data has already been deidentified.

2. **Explain why the waiver of authorization will not adversely affect the rights and welfare of the participants.**
   The identifiers have already been stripped and will not adversely affect the participants.

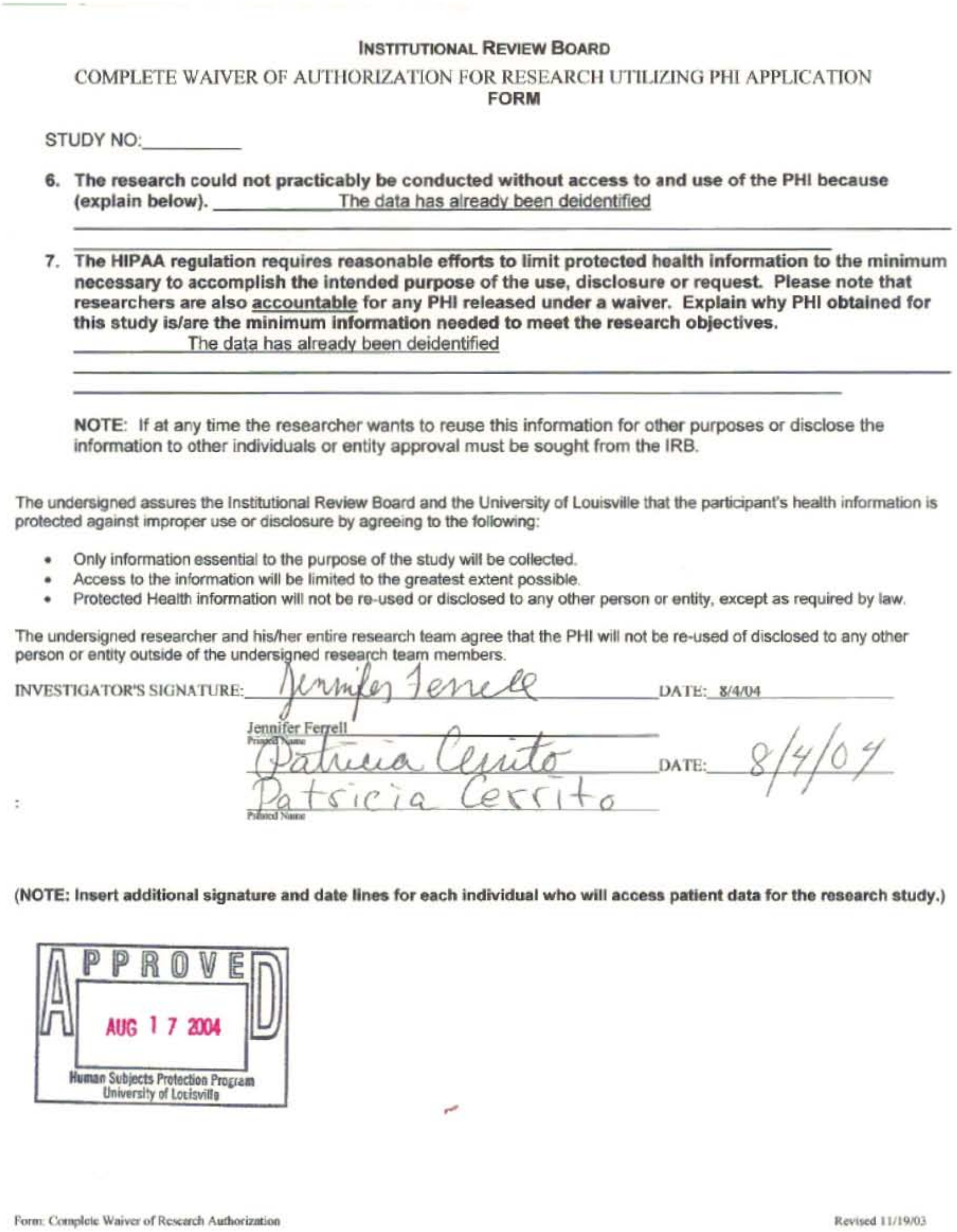
3. **Describe the plan to protect identifiers from improper use and disclosure and indicate where PHI will be stored and who will have access (researchers must list all of the entities that might have access to the study's PHI such as IRB, U of L Audit Services, sponsors, FDA, data safety monitoring boards and any others given authority by law);**
   The identifiers have all been removed.

4. **Select one of the following justifications:**
   All identifiers collected during the study will be destroyed at the earliest opportunity consistent with the conduct of research. Indicate when and how identifiers will be destroyed, OR The identifiers were removed previous to my receiving the data.

   Alternatively, the identifiers collected during the study will not be destroyed because: (explain below).

5. The research could not practicably be conducted without the waiver because (explain below).
   The identifiers have all been removed

Form: Complete Waiver of Research Authorization Revised 11/19/03



INSTITUTIONAL REVIEW BOARD

COMPLETE WAIVER OF AUTHORIZATION FOR RESEARCH UTILIZING PHI APPLICATION
FORM

STUDY NO:__________

6. **The research could not practicably be conducted without access to and use of the PHI because (explain below).** The data has already been deidentified

7. **The HIPAA regulation requires reasonable efforts to limit protected health information to the minimum necessary to accomplish the intended purpose of the use, disclosure or request. Please note that researchers are also accountable for any PHI released under a waiver. Explain why PHI obtained for this study is/are the minimum information needed to meet the research objectives.**
The data has already been deidentified

NOTE: If at any time the researcher wants to reuse this information for other purposes or disclose the information to other individuals or entity approval must be sought from the IRB.

The undersigned assures the Institutional Review Board and the University of Louisville that the participant's health information is protected against improper use or disclosure by agreeing to the following:

- Only information essential to the purpose of the study will be collected.
- Access to the information will be limited to the greatest extent possible.
- Protected Health information will not be re-used or disclosed to any other person or entity, except as required by law.

The undersigned researcher and his/her entire research team agree that the PHI will not be re-used of disclosed to any other person or entity outside of the undersigned research team members.

INVESTIGATOR'S SIGNATURE: Jennifer Ferrell DATE: 8/4/04

Jennifer Ferrell
Printed Name

Patricia Cerrito DATE: 8/4/04

Patricia Cerrito
Printed Name

(NOTE: Insert additional signature and date lines for each individual who will access patient data for the research study.)

APPROVED
AUG 17 2004
Human Subjects Protection Program
University of Louisville

Form: Complete Waiver of Research Authorization

Revised 11/19/03

APPENDIX 2

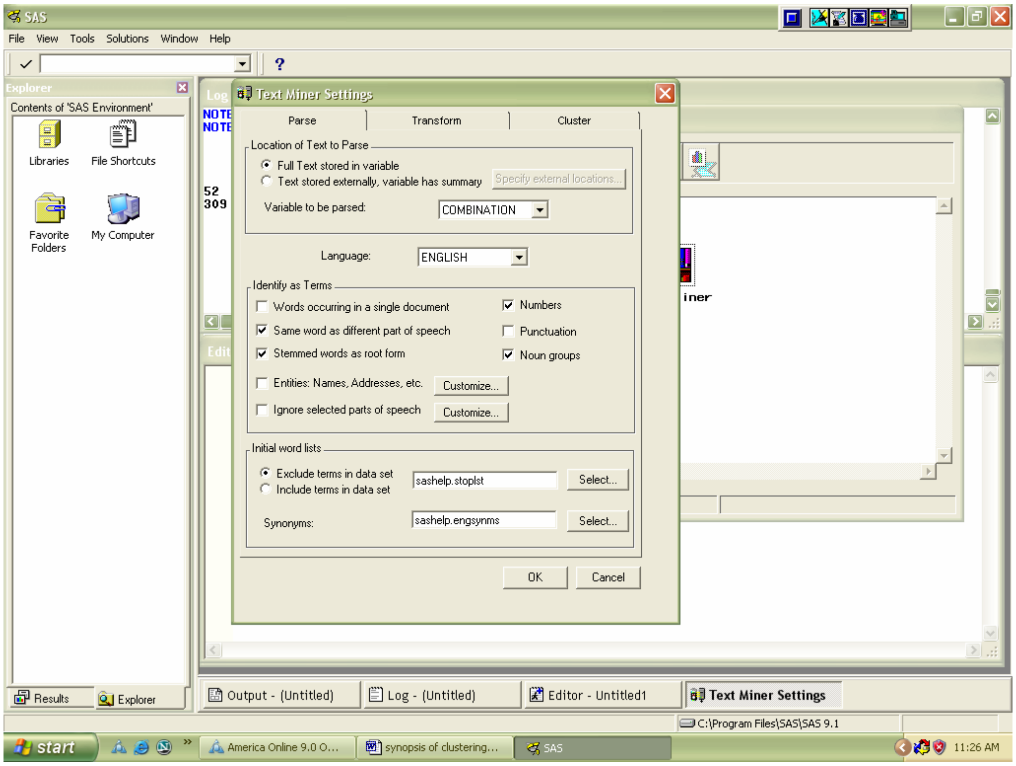
A STEP BY STEP TO TEXT MINING IN SAS

The following is a step by step process of how to cluster the data in the Frazier rehab dataset for analysis:

1. First, we need to create a text string of all of the comorbidity codes for each patient. To do this, in the Excel table, use the following formula: =A2 & " " & B2 & " " & C2 & " " & D2 & " " & E2 & " " & F2 & " " & G2 & " " & H2 & " " & I2 & " " & J2

   This would be the formula for the first person in the dataset. After applying this formula to each patient, this gives a text string of all of the comorbidity codes.
2. Next, we import the Excel table into SAS.
3. Then we go to SAS 9.1 Enterprise Miner. Here, we would like to create a new diagram. To do this, right click on the name in the diagram section. Then, click on the Tools tab at the bottom. Drag “Input Data Source” icon into the white space.
4. To input data, double-click on icon. On the first tab, Data, select the source data you would like to use. This data should be in a SAS library. The other tabs give information on the variables. Nothing was changed on the other tabs. To save, simply click the “x” button. It will ask if you want to save. Say yes.

5. Now we have our data. To cluster this data, I used text miner. In the tools tab, there should be a text miner icon under the explore folder. Drag this icon over to the white space. Placing the cursor over the input data source icon, drag an arrow to connect the two icons.
6. Now double click on the text miner icon. The first tab is parse. See below



Text Miner Settings

Parse | Transform | Cluster

Location of Text to Parse
- ● Full Text stored in variable
- ○ Text stored externally, variable has summary [Specify external locations...]

Variable to be parsed: COMBINATION

Language: ENGLISH

Identify as Terms
- □ Words occurring in a single document
- ☑ Numbers
- ☑ Same word as different part of speech
- □ Punctuation
- ☑ Stemmed words as root form
- ☑ Noun groups
- □ Entities: Names, Addresses, etc. [Customize...]
- □ Ignore selected parts of speech [Customize...]

Initial word lists
- ● Exclude terms in data set
- ○ Include terms in data set

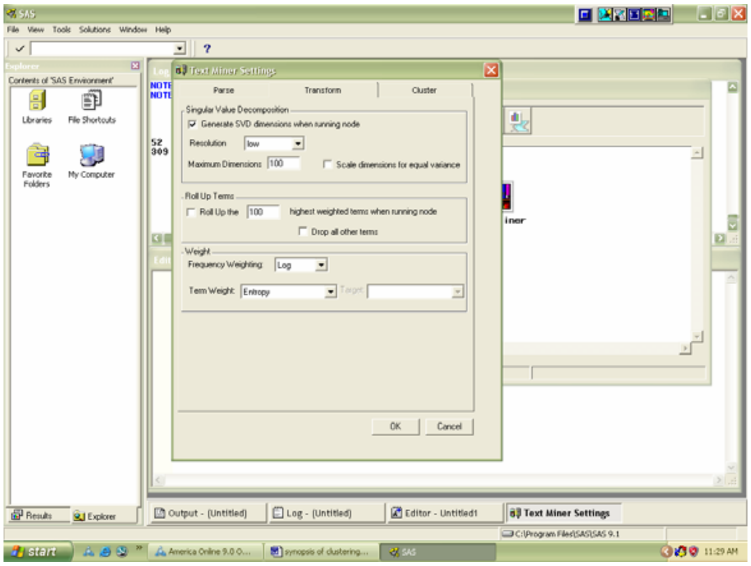
sashelp.stoplst [Select...]

Synonyms: sashelp.engsynms [Select...]

[OK] [Cancel]

This gives options on how to parse the data. Since our comorbidity text strings include numbers, we need to make sure the numbers box is checked. Everything else was left as the default.

7. Now click on the transform tab. See below.



Text Miner Settings

Parse | Transform | Cluster

Singular Value Decomposition

☑ Generate SVD dimensions when running node

Resolution: low

Maximum Dimensions: 100

□ Scale dimensions for equal variance

Roll Up Terms

□ Roll Up the 100 highest weighted terms when running node

□ Drop all other terms

Weight

Frequency Weighting: Log

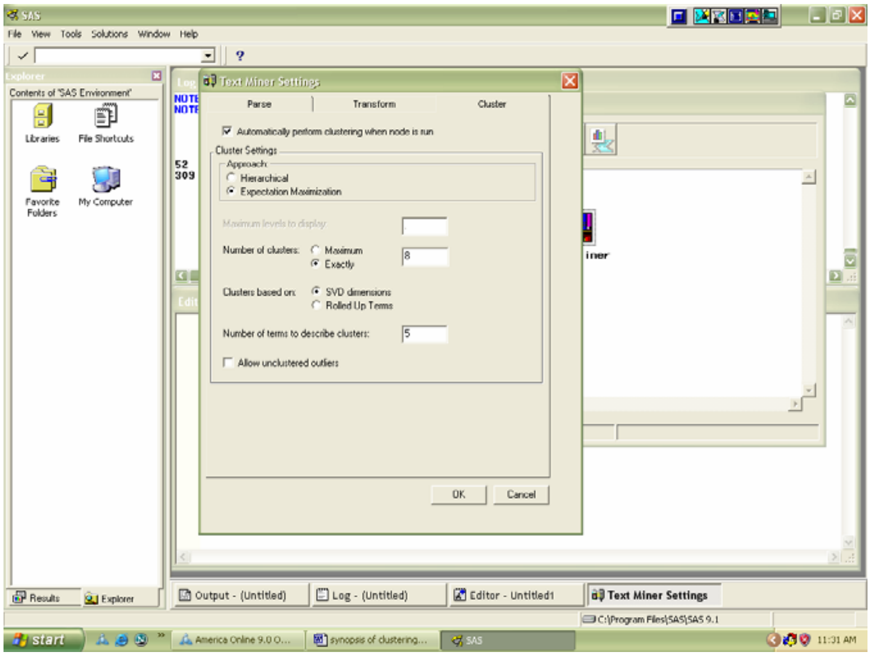
Term Weight: Entropy

Target:

OK | Cancel

This gives the options to change the singular value decomposition terms. I just make sure that the generate button is clicked and leave everything else as the default.

8. Now click the cluster button. This gives the option to cluster the terms automatically.



Text Miner Settings

Parse | Transform | Cluster

☑ Automatically perform clustering when node is run

Cluster Settings

Approach:
○ Hierarchical
● Expectation Maximization

Maximum levels to display:

Number of clusters: ○ Maximum ● Exactly 8

Clusters based on: ● SVD dimensions ○ Rolled Up Terms

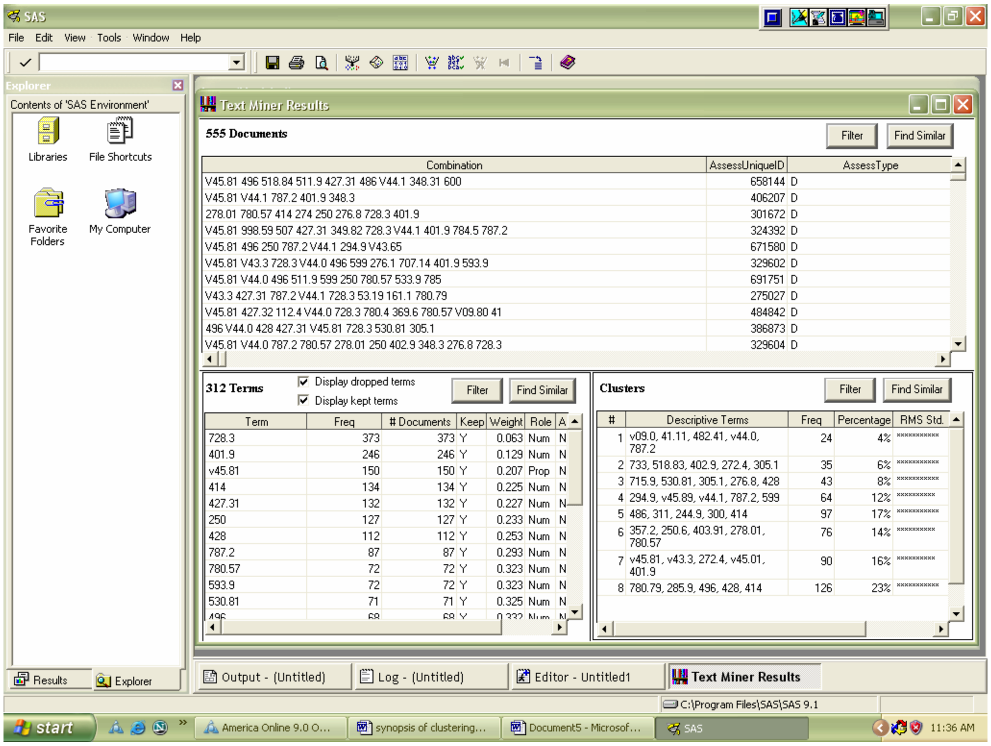
Number of terms to describe clusters: 5

□ Allow unclustered outliers

OK Cancel

First, click on the automatically cluster terms. Then we can specify the number of clusters we would like. Here, I chose 8. We can also specify the number of terms to be displayed. 5 is the default. Descriptive terms are the terms that occur the most frequently in the documents within a cluster. Now click OK at the bottom.

9. To run Text Miner, right click on the icon and click run. This will take a couple of minutes depending on how fast your computer is.
10. After the Text Miner is finished, it will ask if you would like to see the results. Click yes. You will see a screen like the following.



555 Documents

Combination	AssessUniqueID	AssessType
V45.81 496 518.84 511.9 427.31 486 V44.1 348.31 600	658144	D
V45.81 V44.1 787.2 401.9 348.3	406207	D
278.01 780.57 414 274 250 276.8 728.3 401.9	301672	D
V45.81 998.59 507 427.31 349.82 728.3 V44.1 401.9 784.5 787.2	324392	D
V45.81 496 250 787.2 V44.1 294.9 V43.65	671580	D
V45.81 V43.3 728.3 V44.0 496 599 276.1 707.14 401.9 593.9	329602	D
V45.81 V44.0 496 511.9 599 250 780.57 533.9 785	691751	D
V43.3 427.31 787.2 V44.1 728.3 53.19 161.1 780.79	275027	D
V45.81 427.32 112.4 V44.0 728.3 780.4 369.6 780.57 V09.80 41	484842	D
496 V44.0 428 427.31 V45.81 728.3 530.81 305.1	386873	D
V45.81 V44.0 787.2 780.57 278.01 250 402.9 348.3 276.8 728.3	329604	D

312 Terms ☑ Display dropped terms ☑ Display kept terms

Term	Freq	# Documents	Keep	Weight	Role	A
728.3	373	373	Y	0.063	Num	N
401.9	246	246	Y	0.129	Num	N
v45.81	150	150	Y	0.207	Prop	N
414	134	134	Y	0.225	Num	N
427.31	132	132	Y	0.227	Num	N
250	127	127	Y	0.233	Num	N
428	112	112	Y	0.253	Num	N
787.2	87	87	Y	0.293	Num	N
780.57	72	72	Y	0.323	Num	N
593.9	72	72	Y	0.323	Num	N
530.81	71	71	Y	0.325	Num	N
496	68	68	Y	0.332	Num	N

Clusters

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	v09.0, 41.11, 482.41, v44.0, 787.2	24	4%	***********
2	733, 518.83, 402.9, 272.4, 305.1	35	6%	***********
3	715.9, 530.81, 305.1, 276.8, 428	43	8%	***********
4	294.9, v45.89, v44.1, 787.2, 599	64	12%	**********
5	486, 311, 244.9, 300, 414	97	17%	***********
6	357.2, 250.6, 403.91, 278.01, 780.57	76	14%	**********
7	v45.81, v43.3, 272.4, v45.01, 401.9	90	16%	**********
8	780.79, 285.9, 496, 428, 414	126	23%	***********

The top is the dataset with the new cluster column. The Bottom left is the number of terms, in this case comorbidity codes, found in the dataset and their frequency. The bottom right is the number of clusters with the 5 descriptive terms. To save this, right click above the top dataset. Go to file and save documents. This allows you to save directly as a SAS dataset in a SAS library. Specify the library and give a name. This ends the text mining. You can save your diagram before you close down SAS.

11. To use the new dataset, in Enterprise Guide, open a new project and open the SAS dataset.

CURRICULUM VITAE

NAME: Jennifer Ferrell

ADDRESS: 3817 Gregory Lane
Erlanger, KY 41018

DOB: Covington, KY 7/04/1982

EDUCATION
& TRAINING: B.S. Mathematics
University of Louisville
2000-2004

AWARDS: SAS Student Ambassador
SUGI 31 March 2006

"Best Presentation" for Pharmaceutical and Health Sciences
16th Annual Midwest SAS User's Group Conference
Cincinnati, OH October 2005

PUBLICATIONS: "Statistical Analysis of Pacemaker Placement Data"
Undergraduate Journal for the Human Science
http://www.kon.org/urc/urc_research_journal3.html.

"Geographic Masking and Interpretation: A Look at Health Care Providers Proximity to Patients in Jefferson County"
Esri's HealthyGIS newsletter Winter/Spring 2005 edition

"A Comparison of Linear Mixed Models to Generalized Linear Mixed Models: A Look at the Benefits of Physical Rehabilitation in Cardiopulmonary Patients"
16th Annual Midwest SAS User's Group Conference Proceedings

"A Comparison of Proc Mixed and Proc Glimmix"
SUGI 31 Conference Proceedings

NATIONAL
MEETING
PRESENTATIONS: “Statistical Analysis of Pacemaker Placement Data”
University of Louisville Undergraduate Research Symposium
in April 2004

“Statistical Analysis of Pacemaker Placement Data”
Poster and Oral presentation at the Southeastern Section
Meeting of the Mathematical Association of America

“Statistical Analysis of Pacemaker Placement Data”
“Posters at the Capital” day, an annual event sponsored by
The State of Kentucky

“Statistical Analysis of Pacemaker Placement Data”
VIGRE undergraduate math research conference at the Ohio
State University

“Geographic Masking and Interpretation: A Look at Health
Care Providers Proximity to Patients in Jefferson County”
Health GIS conference in Chicago, IL- October 23-26

“A Comparison of Linear Mixed Models to Generalized
Linear Mixed Models: A Look at the Benefits of Physical
Rehabilitation in Cardiopulmonary Patients”
Midwest SAS Users Group meeting in Cincinnati, OH
October 9-11

“A Comparison of Proc Mixed and Proc Glimmix”
SAS® Users Group International (SUGI) 31
March 26-29, 2006