

Paper 161-2011

Sentiment Mining Using SAS® Text Miner®

Patricia Cerrito, University of Louisville

ABSTRACT

In the past, attempts to capture sentiment in studies have used survey questions with Likert scales or qualitative analysis with a very small number of observations. With the use of text analysis provided in SAS Text Miner, it is now possible to use open-ended questions to analyze for sentiment. There are many terms in medical research that have different meanings to different players: patients, physicians, hospitals, payers, and government bureaucracies. With the Affordable Care Act (Obamacare), there will be additional accountability by providers as well as attempts to ration care based upon the definition of the terms, effectiveness, futility, necessary, and reasonable as well as the concept of quality of life. Such attempts are already occurring for the cancer drugs, Avastin and Provenge. Currently, quality of life surveys are based upon one concept developed largely by providers. Quality of life is defined on the basis of functioning only and not upon the basis of social relationships. It is the purpose of this study to examine how open ended questions and SAS Text Miner can be used to investigate how different groups and individuals tend to understand crucial terminology in a practicing discipline and how the term, quality of life, can be better understood.

INTRODUCTION

The literature of medical research is a promising target for text data mining; a large and growing database of medical journal articles or hospital patient records exist in digital format, and the formalized and detailed content delivery style makes them a good subject for computerized text analysis. Because of the large number of journal articles published and large clinical databases, it is unlikely that any one researcher could read, analyze, and organize the contents of all of them. In theory, at least, text mining ought to be able to help the users of these large text databases to find essential semantic characteristics of the available texts and possible linkages even across disciplines.

We want to examine how SAS Text Miner can be used to investigate sentiment when it comes to crucial terminology in a discipline. In this study, we focus on healthcare terms. We demonstrate the means of examining sentiment using open ended survey questions related to the terms, health and quality of life. To a person without a hand, is a hand transplant a benefit? To the surgeons, is a hand transplant of benefit to the patient? A recent study suggests that the two perspectives are extremely different. The majority of surgeons surveyed indicated that it was not beneficial because prosthesis is available without the risk of anti-immune drugs. However, the patients without hands wanted to sign up immediately for a transplant. (Edgell, McCabe et al. 2001)

The patients with metastatic breast cancer will virtually all prefer disease-free survival as a benefit to the use of the drug, Avastin. However, the FDA has decided that disease-free survival is no longer a reasonable benefit; it only counts as a benefit if there is overall survival. Such changing meaning to basic terminology can drastically change results and patient treatment options. (Anonymous-WSJ 2010)

We would like to ensure that all of the terminology used in defining models is both reliable and valid. However, if different groups view the terminology with different concepts and sentiments, then the validity is questionable. In comparative effectiveness analysis, the concepts that we need to focus on are health, perfect health, and quality of life. Additional terms that should be considered are disability, utility, extraordinary medical care, palliative care, and comfort care. Perfect health is hard, if not impossible, to define. Some argue that there are health states worse than death, and that there should be negative values possible on the health spectrum. Determining the level of health depends on measures that some argue place disproportionate importance on physical pain or disability over mental and emotional health. The effects of a patient's health on the quality of life of others (e.g. caregivers or family) do not figure into these calculations.

In yet another aspect of sentiment mining, we can investigate patient feedback concerning their use of medications with severe side effects. We can also see if patients find the severity and the corresponding diminished quality of life too great to continue treatment. We will do this while investigating the drug, Neulasta, which has known side effects of severe pain (in over half the treatment population) as well as the potential for a ruptured spleen. While such patients may have a lowered quality of life, almost all of them are willing to tolerate it in order to improve their health.

THE USE OF COMPARATIVE EFFECTIVENESS ANALYSIS

The National Health Service in Britain has been using comparative effectiveness analysis for quite some time. NICE stands for the National Institute for Health and Clinical Excellence; this organization has defined an upper limit on treatment costs, and if the adjusted cost exceeds this pre-set limit, then the treatment is denied. It does not matter if

the drug is effective or not; it does not matter if the unadjusted cost is within the threshold. That means that there are many beneficial drugs that are simply not available to patients in Britain; fully 25% of cancer patients are denied effective chemotherapy medications.

NICE is not comparing drug A to drug B. Instead, the organization compares the cost of a drug to the value the organization places on your life. If it costs too much to keep you alive given your value, or to improve your life, then you are denied treatment. Oregon has become notorious in its Medicaid benefit, denying cancer drugs to patients, but making the same patients aware that assisted suicide is available. Oregon will not make available drugs that can prolong a patient's life; it will make available a drug to end it (which will then save additional medical costs). (Springer 2008) Currently, pharmaceutical companies have been subsidizing Oregon's Medicaid by providing these drugs to patients who have been denied by Medicaid. (Smith 2009) It has been suggested that euthanasia is cheaper than end of life care, and more cost-effective than treating many patients with terminal illnesses. (Sprague 2009) Just recently, the Food and Drug Administration has considered retracting approval of a chemotherapy drug for breast cancer on the basis of cost effectiveness rather than effectiveness.

A comparative effective analysis starts with the perceived patient's utility given the disease burden. The QALY, or quality of life-adjusted years is an estimate of the number of years of life gained given the proposed intervention. Each year of perfect health is assigned a value of 1.0. A patient in a wheelchair is given a correspondingly lower value as is a patient who is elderly; this value is not clearly defined and is rarely based upon patient input.

Consider an example. Suppose a cancer drug for patients with liver cancer allows a patient to live an average of 18 months compared to not using the drug. However, as with most cancer drugs, there are potent side effects. Suppose that the analyst decides that the quality of life is only 40% of perfect health (giving a weight of 0.4). Then, the drug gives $1.5 \times 0.4 = 0.6$ QALYs to the patient. Suppose that at the initial introduction of this drug, it costs \$1000 per month, or about \$18,000 for the anticipated additional life of the patient. Then the cost per QALY is equal to $18,000 / 0.6 = \$30,000$ per year of life saved. According to the NICE organization, this drug then can become too costly regardless of the fact that there is no comparable drug that is effective in prolonging the patient's life. However, suppose the analyst uses a measure of 60% of perfect health. Then the drug gives $1.5 \times 0.6 = 0.9$ QALYs to the patient at a cost of \$20,000, which brings the amount closer to the pre-set value defined by NICE. Therefore, this definition of a scale of perfect health is of enormous importance. In fact, NICE has denied such a cancer drug because of its cost.

If a person is otherwise young and healthy and a drug costs \$10,000 per year, then the QALY is \$10,000. However, if a patient is older and has a chronic condition, then that patient's utility may be defined as exactly half that of a young and otherwise healthy person. In that case, the QALY is \$20,000 for the same drug. If the patient is old and has two or more chronic conditions, then the patient's utility could be defined as 25% that of a young and healthy person. In that case, the QALY IS \$40,000 per year of life saved. If the organization defines \$30,000 as the upper limit for treatment, it is easy to see how the definition of a person's utility can be used to deny care to the elderly.

However, the cost of treating the disease is not restricted to the cost of medications. Therefore, we must look at all aspects of treatment, including physician visits, hospital care, and home health care. We must also look at the impact of patient compliance on the overall cost of healthcare. If patients have specific diseases that can be treated, but who do not use the treatment, then outcomes will not be the same compared to patients who do comply. Also, patients who switch treatments may suffer from adverse events of the first treatment that are not present in the second treatment. Therefore, we must examine the totality of patient care.

METHOD

Text mining is a variation on the field of data mining that tries to find interesting patterns from large databases. The patterns discovered provide information that can be extracted to derive summaries of the words contained in the documents, or to compute summaries for the documents based on the words contained in them. Hence, the investigator can analyze words or clusters of words used in documents, or to analyze entire documents to determine similarities or relationships between them, or how they are related to other variables of interest in the data-mining project. In the most general terms, text mining will "turn text into numbers".

There has been tremendous growth in the volume of online text documents available on the Internet, digital libraries, news sources, and company-wide intranet. These documents (with other unstructured data) will become the predominant data type stored online. The growing importance of online documents has led to a great interest in tools and approaches for dealing with unstructured or semi-structured information stored in the text documents. The possibilities for data mining from large text collections are virtually untapped. Text can provide very complex information, but it also encodes this information in a form that is difficult to decipher automatically. For this reason, there have been few statistical applications involving text mining to date, other than to use it to flag terms.

Generally, a document is converted into a row in a matrix. This row has a column for any word contained within the dataset of documents. The matrix value is equal to the number of times that the word occurs in the document. The matrix will consist mostly of zeros since the list of words is much longer than the list of documents. Therefore, the next step is to reduce the dimension of the matrix. This is done through the process of singular value decomposition (SVD). This feature is extremely valuable for calls into customer service, for example, for chat notes, and to examine advertisements from the competition.

There are variations to this general methodology depending upon what you want to discover. For example, if you want to determine what documents contain a specific word for flagging purposes, this can be done through filtering. Otherwise, the SVD of an $N \times p$ matrix A having N documents and p terms is equal to $A=U\Sigma V$ where U and V are $N \times p$ and $p \times p$ orthogonal matrices respectively. U is the matrix of term vectors and V is the matrix of document vectors; Σ is a $p \times p$ diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$, called the singular values of Σ . The truncated decomposition of A is when SVD calculates only the first K columns of U , Σ and V . SVD is the best least squares fit to A . Each column (or document) in A can be projected onto the first K columns of U . Similarly, each row (or term) in A can be projected onto the first K columns of V . The columns projection (document projection) of A is a method to represent each document by K distinct concepts. So, for any collection of documents, SVD forms a K dimensional subspace that is a best fit to describe data.

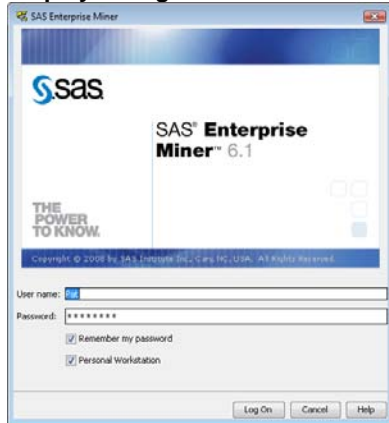
The next step is to cluster the documents. Cluster analysis is used to form descriptive statistics to assess whether or not the data consist of a set of distinct subgroups; each subgroup representing objects with substantially different properties. Central to all of the goals of cluster analysis is the notion of the degree of similarity or dissimilarity between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it. Clustering algorithms fall into three distinct types: combinatorial algorithms, mixture modeling and mode seeking. Text Miner has as its basis the Expectation Maximization Algorithm. The expectation maximization (*EM*) algorithm uses a different approach to clustering in two important ways:

1. The *EM* clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm is to maximize the overall probability or likelihood of the data, given the final clusters.
2. Unlike k-means clustering, the general *EM* algorithm can be applied to both continuous and categorical variables.

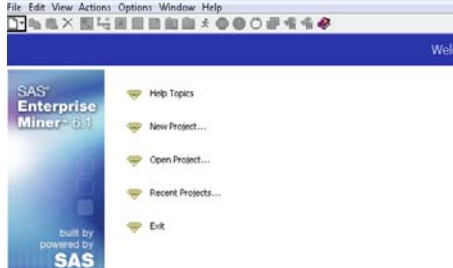
The expectation-maximization algorithm is used to estimate the probability density of a given set of data. *EM* is a statistical model that makes use of the finite Gaussian mixtures model and is a popular tool for simplifying difficult maximum likelihood problems. The algorithm is similar to the K-means procedure in that a set of parameters is re-computed until a desired convergence value is achieved. The finite mixture model assumes all attributes to be independent random variables.

INITIATING A PROJECT IN SAS ENTERPRISE MINER

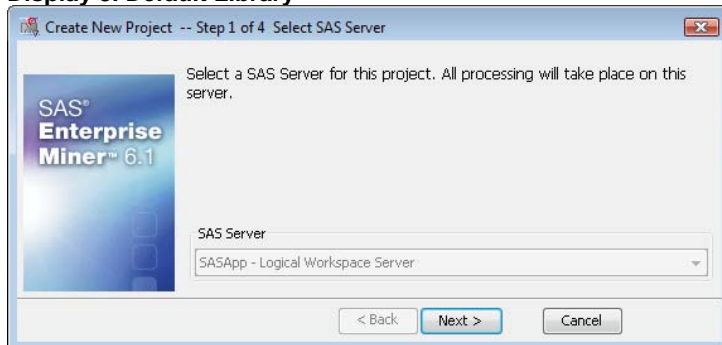
We first examine the initial screens in Enterprise Miner that are used to set up a text analysis project. The software component gives flexibility in terms of setup and storage so that you can identify the location of the project and the corresponding data sets. The software uses a java interface, allowing you to use point-and-click access once the project has been set up. Display 1 shows the beginning screen in SAS Enterprise Miner.

Display 1. Login Screen for SAS Enterprise Miner

Display 2 gives the next screen so that you can indicate whether a new project will be started, or a previous project continued. For the first time, you should check for a new project.

Display 2. Second Enterprise Miner Screen

Once you choose new project, a default library is shown (Display 3). Use this default unless you have a good reason to change it. If only one server is available, the value will be grayed out and you will have to use the default server.

Display 3. Default Library

You need to give a name to the project (Display 4). A default location is given for storing the project. You can change the path, if necessary. Also, you need to define a path to store the project. Just use a path to a directory on your hard drive.

Display 4. Project Name

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name

SAS Server Directory

< Back Next > Cancel

To complete the definition of a project, Displays 5 and 6 give summary information concerning that project. You can browse for a location.

Display 5. Step 3 of Creating a Project

Select the SAS Folders location for this project. Use these folders to organize your projects and control user access.

SAS Folder Location

/Users/sastrust/My Folder

< Back Next > Cancel

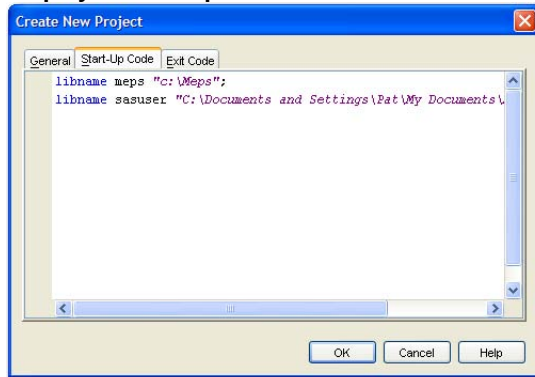
Display 6. Summary of Information

New Project Information

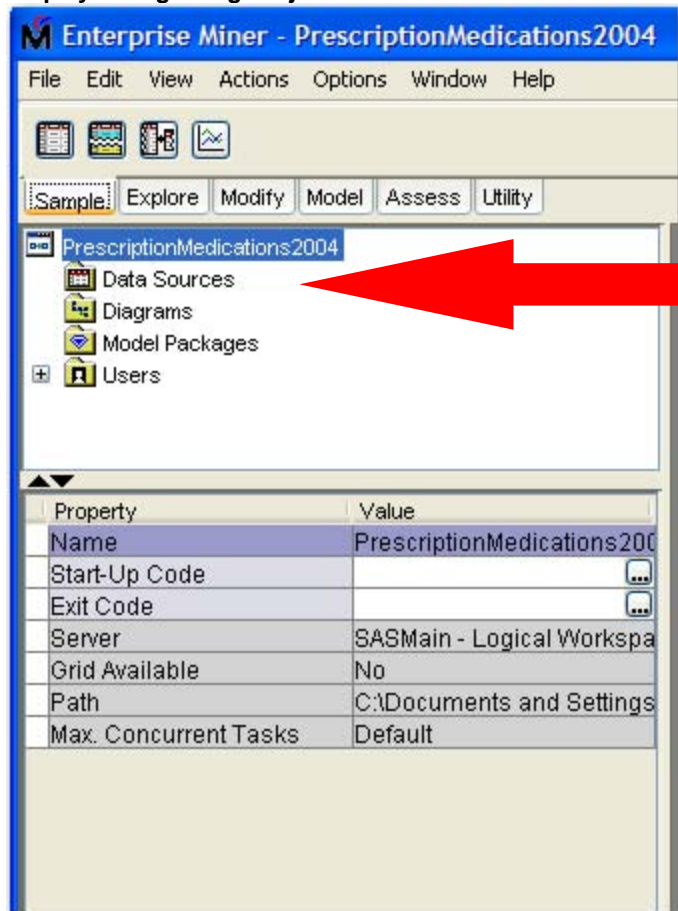
Name	test
SAS Metadata location	/Users/sastrust/My Folder
SAS Application server	SASApp - Logical Workspace Server
Server Directory	H:\

< Back Finish Cancel

The next step to start a new project is to click on the Start-Up Code tab. You need to define a library name to store datasets for the project. Enterprise Miner does not automatically use the Sasuser library. You should probably define the SASUSER directory with a Libname statement in case you need to access files from this library that were defined by other SAS components. The code needed is given in Display 7.

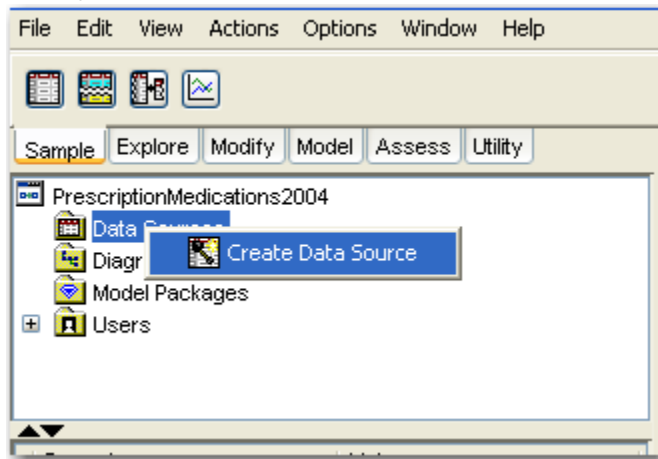
Display 7. Start-Up Code

Once this step is completed, the project screen is displayed (Display 8). There are slots available for datasets and diagrams; both must be created in order to develop a project. Right-clicking on Data Source will allow you to enter a dataset (Display 9).

Display 8. Beginning Project Screen

Right-click on Data Source

Once you click on data source, you will be prompted to create a data source for the project.

Display 9. Create Data Source

A Data Source Wizard comes up to aid you in entering a dataset (Display 10). Just follow the steps to bring a dataset into the project.

Display 10. Data Source Wizard

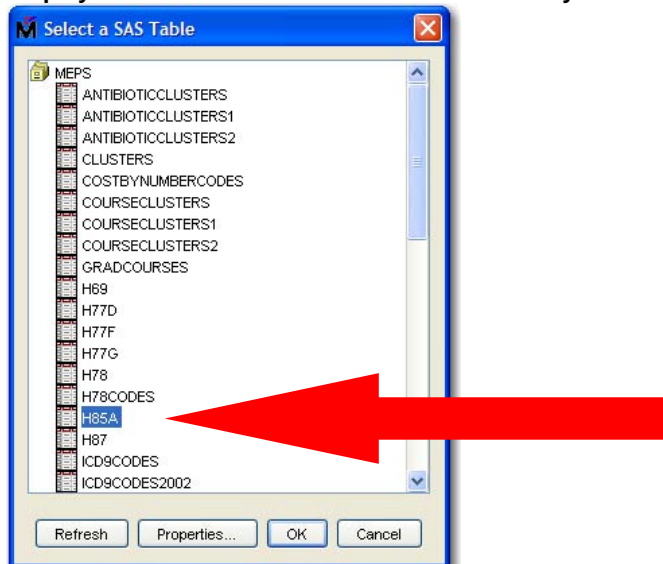
The next screen allows you to browse for a SAS dataset stored on your harddrive (Display 11). The available SAS libraries are listed (Display 12).

Display 11. Second Data Screen

The Sampsio and Sashelp libraries are always available. Additional libraries were defined in the Start-up Code using the libname statement.

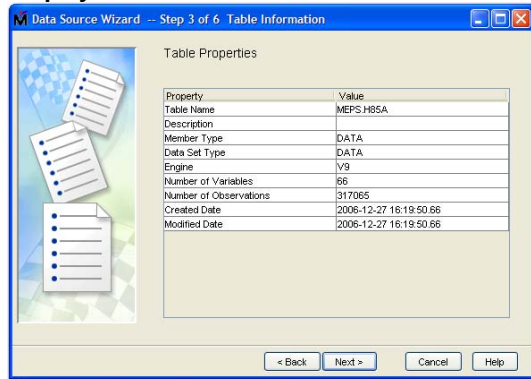
Display 12. Available Libraries

Choose one of the libraries. If the library is not shown, or if the desired dataset does not appear in the library, use the refresh button. It should appear after that. If it does not, check the Start-up code to make certain that it was entered correctly. Display 13 shows the available datasets. The next screen (Display 14) shows some of the metadata; that is, information concerning the dataset. It gives the number of variables and the number of observations. The dataset from MEPS, HC-85A contains 317,065 observations.

Display 13. Available Datasets in the MEPS library

Display 14 allows you to compare the information in the metadata server to the information that you have concerning the data set.

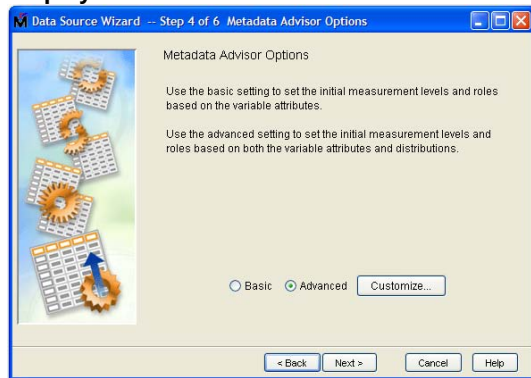
Display 14. Metadata About the Data



Display 15 gives the option of the basic or advanced displays concerning the data variables. We recommend that the advanced option is used. You can change any of the default roles or levels (Displays 16 and 17). The role is changed by clicking the mouse at the right location.

This display gives some basic information about the dataset, including the number of variables, the number of observations, and the date the dataset was last modified.

Display 15. Choice of Basic or Advanced



You can choose Basic or Advanced. Advanced gives more choices for level (for example, ordinal) compared to basic. Level defines the type of variable. It can be nominal, ordinal, or interval, but also binary.

Display 16. Information About the Variables in the Dataset

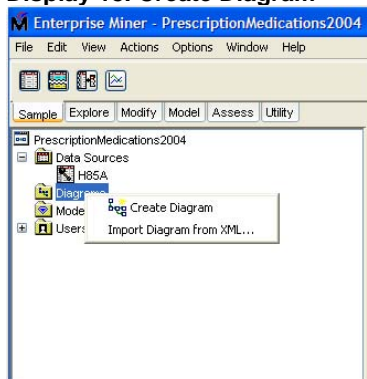
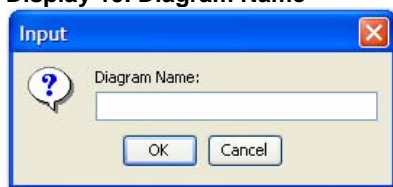
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Type	Label	Fr
CLMCHFLD	Input	Nominal	No					N	CHOSEPWMT_RV CLAIM FLNO, OMTYPE STAT	
CLSD	Input	Interval	No					N	INWELLING UNIT ID	
CONFREQD	Projected	Nominal	No					C	PERSON IS GUID + FID	
INPCFLG	Input	Binary	No					N	FID HAS AT LEAST 1 RECORD IN PC	
LINKDC	Projected	Nominal	No					C	IS FOR LINKAGE TO CONDOH EVENT FILE	
PCIMPFLD	Input	Binary	No					N	TYPE OF HC TO PC PRESCRIPTION MATCH	
PERWT04T	Input	Interval	No					N	FINAL PERSON LEVEL WEIGHT_2004	
PHARTF1	Input	Nominal	No					N	TYPE OF PHARMACY FROW_1ST	
PHARTF2	Input	Nominal	No					N	TYPE OF PHARMACY FROW_2ND	
PHARTF3	Input	Nominal	No					N	TYPE OF PHARMACY FROW_3RD	
PHARTF4	Input	Nominal	No					N	TYPE OF PHARMACY FROW_4TH	
PHARTF5	Input	Nominal	No					N	TYPE OF PHARMACY FROW_5TH	
PHARTF6	Input	Nominal	No					N	TYPE OF PHARMACY FROW_6TH	
PHARTF7	Input	Nominal	No					N	TYPE OF PHARMACY FROW_7TH	
PID	Input	Interval	No					N	PERSON NUMBER	
PRESGAT	Input	Nominal	No					C	MULTI-PREGNANCY CATEGORY	
PURCHASD	Input	Nominal	No					N	ROUND NUMBER FOR MED OBTAINED/PURCHASED	
ROBESDCC	Input	Interval	No					N	DAY PERSON STARTED TAKING MEDICINE	
ROBESDMM	Input	Nominal	No					N	MONTH PERSON STARTED TAKING MEDICINE	
ROBESDYS	Input	Interval	No					N	YEAR PERSON STARTED TAKING MEDICINE	
ROSCC1A	Projected	Nominal	No					C	MODIFIED CLINICAL CLASS CODE	
ROSCC2A	Projected	Nominal	No					C	MODIFIED CLINICAL CLASS CODE	
ROSCC3A	Projected	Nominal	No					C	MODIFIED CLINICAL CLASS CODE	
ROFLG	Input	Nominal	No					N	NEC IMPUTATION SOURCE ON PC DONOR REC	
ROFORM	Projected	Nominal	No					C	FORM OF RHPRES-CRIBED MEDICINE (IMPUTED)	
ROFORMAT	Input	Nominal	No					C	UNIT OF MEAS FORM RHPRES-CRIBED MED (IMPUTED)	
ROHNAME	Projected	Nominal	No					C	HC REPORTED MEDICATION NAME	
ROSCD1A	Projected	Nominal	No					C	3 DIGIT ICD-9 CONDITION CODE	
ROSCD2A	Projected	Nominal	No					C	3 DIGIT ICD-9 CONDITION CODE	
ROSCD3A	Projected	Nominal	No					C	3 DIGIT ICD-9 CONDITION CODE	
ROMSD4A	Input	Interval	No					N	AMOUNT PAID, MEDICARE (IMPUTED)	
ROMSD4B	Input	Interval	No					N	AMOUNT PAID, MEDICARE (IMPUTED)	
ROMSD4C	Input	Interval	No					N	AMOUNT PAID, OTHER FEDERAL (IMPUTED)	
ROMSD4D	Input	Interval	No					N	AMOUNT PAID, OTHER FEDERAL (IMPUTED)	
ROMSD4E	Input	Interval	No					N	AMOUNT PAID, OTHER PRIVATE (IMPUTED)	
ROMSD4F	Input	Interval	No					N	AMOUNT PAID, OTHER INSURANCE (IMPUTED)	

Display 17. Changing Variable Role

Variable	Role	Level	Target
PHARTP7	Input	Nominal	No
PID	Input	Interval	No
PREGCAT	Input	Nominal	No
PURCHRDR	Input	Nominal	No
RXBEGDD	Label	Interval	No
RXBEGMM	Prediction	Nominal	No
RXBEGYRX	Referrer	Interval	No
RXCCC1X	Rejected	Nominal	No
RXCCC2X	Residual	Nominal	No
	Segment		
	Sequence		

There are roles that include sequence, target, and text. Sequence is sometimes needed to list entries in order. Target is for an outcome variable. Text can be used for nominal data; it must be used with the Text Miner node. Although Text Miner does not examine numbers by default, that default can be changed so that it will work with nominal data defined in terms of numbers. The next screen prompts you for the role of the dataset. The default role is raw. You can change the role when needed for other applications in Enterprise Miner.

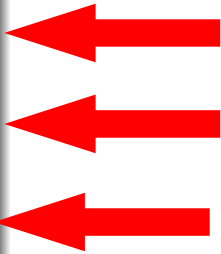
Once the needed datasets have been entered into the project, the next step is to define a diagram (Display 18). Right click on the diagram menu item. You will be prompted for a Diagram name (Display 19). You can have multiple diagrams for a project just as you can multiple data sets available to set up in any one diagram.

Display 18. Create Diagram**Display 19. Diagram Name**

Once the diagram is named, a diagram window is opened on the right side of this menu. In the diagram, you drag-and-drop icons from the menu at the top of the Enterprise Miner screen. Once the diagram is created, you can drag and drop the data set into the diagram along with the Text Miner icon. Then use the mouse to connect the two icons. Click on the Text Miner node to highlight it. When highlighted, the defaults for the icon are listed on the left-hand side of the display. We want to look at the defaults for Text Miner (Display 20).

Display 20. Defaults in Text Miner

Node ID	TEXT3
Imported Data	
Exported Data	
Variables	
Interactive	
Rerun	No
Parse	
Parse Variable	Diagnoses
Language	ENGLISH
Stop List	
Start List	DIABETES.DIABETESST
Stem Terms	Yes
Terms in Single Document	No
Punctuation	No
Numbers	Yes
Different Parts of Speech	No
Ignore Parts of Speech	
Noun Groups	No
Synonyms	SASHELP.ENGSYNMS
Find Entities	No
Types of Entities	
Transform	
Compute SVD	No
SVD Resolution	Low
Max SVD Dimensions	100
Scale SVD Dimensions	No
Frequency weighting	Log
Term Weight	None
Roll up Terms	No
No. of Rolled-up Terms	20
Drop Other Terms	No



Usually, we find that it is better not to use different parts of speech. It is also helpful to set up a start list eliminating unnecessary terms, and to identify some synonyms that have essentially the same meaning that you want to keep is equivalent.

RESULTS

We give the results of two analyses involving sentiment. The first has to do with a survey of open-ended questions related to the quality of life while the second study was performed on a collection of comments and messages related to the use of a potent medication with very severe side effects. In both studies, we use SAS Text Miner to analyze results.

Definition of Health and Quality of Life

For the first study, we requested that 85 nursing students define the terms, health, perfect health, and quality of life. We entered these definitions into a SAS dataset and then used SAS Text Miner. It is possible to ask different players (providers, payers, patients, and government officials) to define these terms in a similar fashion to compare clusters of results, and to see if the terms are regarded in a similar fashion, or if different players define the terms differently. Table 1 shows the resulting clusters from Text Miner for the definitions of perfect health.

Table 1. Clusters of Definitions for the Term, Perfect Health

Clusters				
#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1	+ condition, mental, medical, + problem, + illness	21	0.2625	0.1859689...
2	+ absence, human body, human, within, + ability	11	0.1375	0.1444446...
3	perfect health, perfect, + do, + not, health	19	0.2375	0.1883746...
4	when, diet, with, without, in	16	0.2	0.2072777...
5	stable, mentally, well, state, free	13	0.1625	0.1922998...

In two of the five clusters, the emphasis is on the absence of disease; in another two clusters, the emphasis is on a good mental condition with the last cluster focused on diet. Other players could very well cluster differently to focus on relationships rather than function. Table 2 gives the clusters for the definition of quality of life.

Table 2. Clusters of Definitions for the Term, Quality of Life

Clusters				
#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1		42	0.525	0.0
2	physically, health, emotionally	3	0.0375	0.1560132...
3	happy, ability, own, one, life	11	0.1375	0.1966319...
4	bad, good, + will, life	3	0.0375	0.1163955...
5	+ enjoy, in, alive, + thing, + not	7	0.0875	0.2288251...
6	need, + do, + can, on, patient	8	0.1	0.2202256...
7	+ disease, independent, healthy, life	3	0.0375	0.2080321...
8	physical, with, well, emotional, able	3	0.0375	0.1648456...

Two of the clusters concentrate on being happy and enjoying life. While physical ability is important, it is combined with emotional ability in those clusters. One of the clusters with just 3 members concentrates on the absence of disease and the ability to lead an independent, healthy life. Table 3 gives the clusters for the term, health.

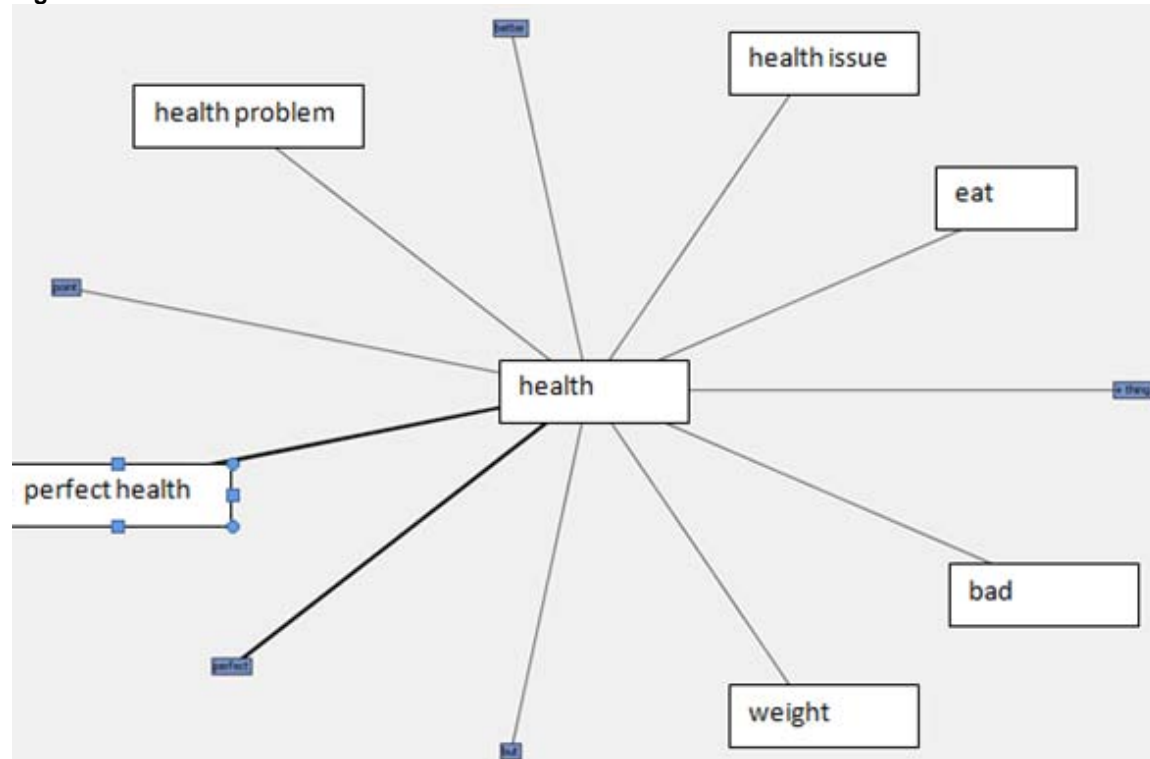
Table 3. Clusters of Definitions for the Term, Health

Clusters				
#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1	disease, free, exercise, medical, eat	12	0.15	0.1975094...
2	state, overall, individual, physical, well	24	0.3	0.1746901...
3	problem, healthy, not, living, health	8	0.1	0.1815526...
4	feeling, good diet, have, diet, lifestyle	3	0.0375	0.1714232...
5	life, mentally, physically, do, basis	33	0.4125	0.1825559...

As it turns out, the concept of health and that of perfect health are defined in very similar fashions. The nursing students were unable to really distinguish between the two terms. Another feature of SAS Text Miner is to use the visualization of concept links to see how terms are linked in these short definitions. Figure 1 shows the terms related to perfect health.

The links are related to the ability to perform tasks. Similarly, the words linked to health include the concept of perfect health, which again shows that it is difficult to distinguish between the definitions of health and perfect health.

Figure 3. Terms Related to Health



Sentiment Regarding Side Effects of Medications

In this study, we used a total of 70 messages related to the use of the drug, Neulasta. The drug has very potent side effects (listed at <http://www.rxlist.com/neulasta-drug.htm>). They include severe bone and muscle pain and the possibility of a ruptured spleen. The purpose of the medication is to grow white cells in cancer patients who have had white cells depleted because of chemotherapy treatments in order to prevent infections.

The messages were written by patients, or by caregivers of patients regarding their personal experiences. The messages are relatively short, but are related to quality of life. Interestingly, only two of the messages contained the word, "stop", indicating that most patients thought that the medication was worthwhile in spite of the problem of pain. As before, we first examine the clusters to see the most important terms involved in each group (Table 4).

Table 4. Clusters of Messages Regarding Neulasta

Cluster #	Description	Frequency	Percentage	Label
1	+ injection, + agree, + side effect, severe, side, + hurt, but, + drug, + effect, only, down, + experience, chemo, + good, + round, + make, with, + do, + day, in	34	0.4927	Severe pain
2	bring, case, + week, up, + keep, same, + receive, back, during, much, + leg, little, blood, ever, + would, body, + count, + shot, + give, as	7	0.1014	Back pain
3	+ eat, before, + call, walk, no, into, + hour, + come, + doctor, when, chemo, + treatment, + bad, + start, + help, now, + give, + time, with, can	25	0.3623	Moderate effects
4	place, + cell, white, even, may, out, + make, + want, + know, + drug, can, up, + week, + would, + do, + count, + feel, + give, like, + have	2	0.02898	Reason for injection

Cluster #	Description	Frequency	Percentage	Label
5	+ week, + bone, + not, + have	1	0.01449	Bone pain

Note that we have added tentative labels to each of the clusters. Labels are extremely useful when attempting to identify the most important relations in the clusters. Almost half of the messages are in cluster 1, which is characterized by the term, "severe". Another 36% indicate moderate pain. These messages give a strong sentiment that those who face difficult alternatives opt in favor of treatment. Table 5 is filtered to those messages that contain the word, "pain".

Table 5. Messages From Table 4 Restricted to "Pain"

1	+ injection, + agree, + side effect, severe, side, + hurt, but, + drug, + effect, only, down, + experience, chemo, + good, + round, + make, with, + do, + day, in	15	0.4634
2	bring, case, + week, up, + keep, same, + receive, back, during, much, + leg, little, blood, ever, + would, body, + count, + shot, + give, as	5	0.1219
3	+ eat, before, + call, walk, no, into, + hour, + come, + doctor, when, chemo, + treatment, + bad, + start, + help, now, + give, + time, with, can	16	0.3902
4	place, + cell, white, even, may, out, + make, + want, + know, + drug, can, up, + week, + would, + do, + count, + feel, + give, like, + have	0	0
5	+ week, + bone, + not, + have	1	0.02439

More than half of the original messages contain the word, pain. Of that number, most (85%) are in cluster 1, indicating severe pain and in cluster 3, indicating moderate pain. We again look at concept links to see how words are connected in the messages. Figure 4 gives the words connected to pain. It indicates that the pain can cause discomfort, or that it can be severe. These connections are quite similar to the different levels of pain identified in clusters 1 and 3.

Figure 4. Words Connected to Pain

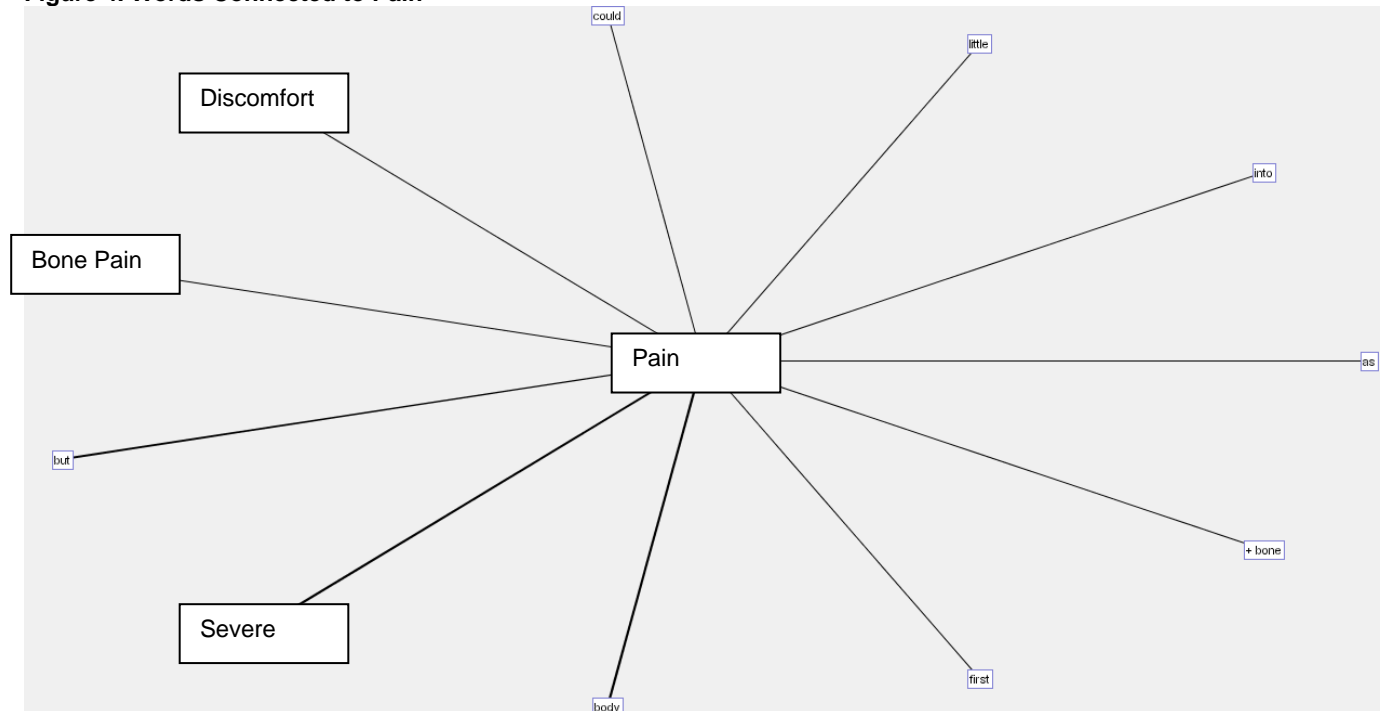
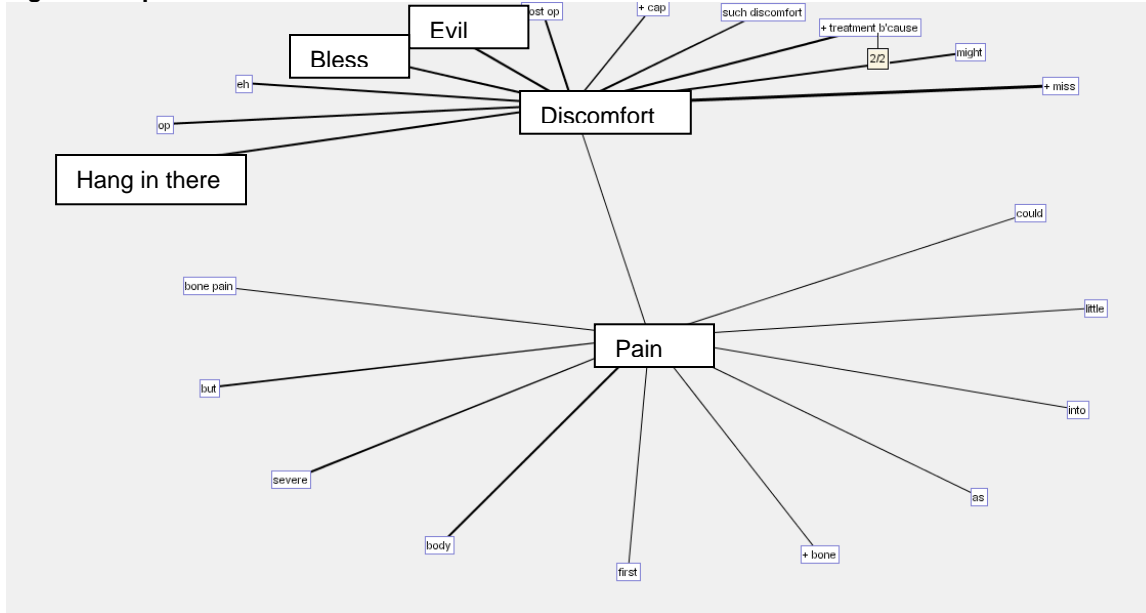


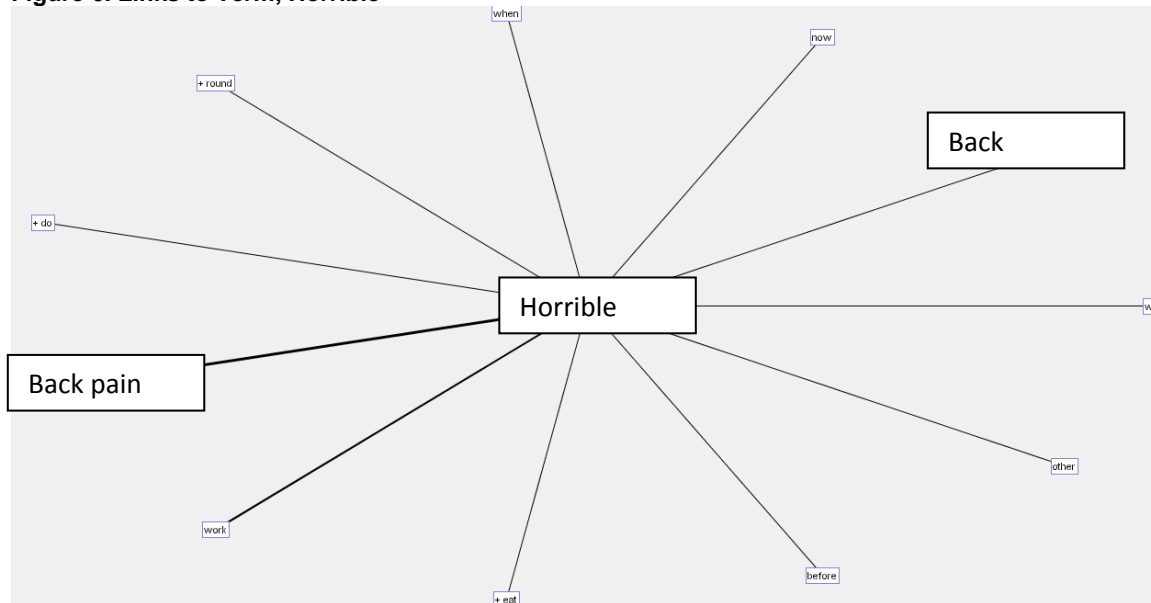
Figure 5 gives an expanded link to see what is connected to both pain and discomfort.

Figure 5. Expanded Links to Pain and Discomfort



Two of the terms relate to encouragement to continue in treatment, likely because the pain will stop once treatment with the drug is discontinued. Figure 6 examines links to the term, horrible, which indicates a severity that is greater than just discomfort. It indicates that horrible is connected to back and back pain, both of which indicate that the severity is in the back.

Figure 6. Links to Term, Horrible



DISCUSSION

SAS Text Miner allows us to investigate the sentiment related to specific, important, but vague concepts that can lead to important results concerning comparative effectiveness analysis. We can look at clusters, which can identify some sentiments, or we can look at some specific terms that indicate sentiment and see how they are related to other terms in the document. Sentiment mining is particularly useful when identifying terms that relate to a patient's quality of life, and how patients try to cope with disease and with treatments that can affect that quality of life. Since comparative effectiveness analysis uses quality of life to define the adjusted cost of treatment, we need to move beyond just a simple definition of function to look at all aspects of quality. Moreover, we need to investigate how different

participants in healthcare examine quality of life: insurers, physicians, hospitals, and patients. If there are differences in the definition of quality, it is extremely important to include the patient's perspective on quality in any comparative analysis.

REFERENCES

1. Anonymous-WSJ (2010). The Avastin Mugging: The FDA rigs the verdict against a good cancer The Wall Street Journal. New York, WSJ.com. **August 18, 2010**.
2. Edgell, S., S. McCabe, et al. (2001). "Different reference frames can lead to different hand transplantation decisions by patients and physicians." Journal of Hand Surgery **26**(2): 196-200.
3. Smith, W. J. (2009) Save money by killing the sick: euthanasia as health care cost containment not such a parody as the author may think.
4. Sprague, C. (2009) The economic argument for euthanasia.
5. Springer, D. (2008). Oregon offers terminal patients doctor-assisted suicide instead of medical care. Fox News. New York, Fox News. **July 28, 2008**.

ABOUT THE AUTHOR



Patricia Cerrito, PhD, Professor of Mathematics at the University of Louisville, has spent over 20 years investigating health outcomes using data mining tools. Dr. Cerrito has been recognized as one of America's Elite Educators. She has published a number of books on the general topic including

[Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks](#)
[Text Mining Techniques for Healthcare Provider Quality Determination: Methods for Rank Comparisons](#)
[Clinical Data Mining for Physician Decision Making and Investigating Health Outcomes: Methods for Prediction and Analysis](#)

CONTACT INFORMATION

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

Name	Patricia Cerrito	
Enterprise	University of Louisville	
Address	Department of Mathematics	
City, State ZIP	40292	
Phone:	502-852-6826	502-742-0889
Fax:	502-852-7132	
E-mail:	pcerrito@gmail.com	
Web site:	drpatriciacerrito.com	

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