Using SAS Predictive Modeling to Investigate the Asthma's Patient Future Hospitalization Risk

Yehia H. Khalil, University of Louisville, Louisville, KY, US

ABSTRACT

The focus of this study is to develop predictive models to forecast future Asthma hospitalizations for patients diagnosed with Asthma. Using SAS® Software tools such as predictive models provides great benefit to insurance companies, hospitals, and pharmaceutical companies trying to manage their resources and define future plans and policies. The strength areas of predictive modeling are: a) an ability to incorporate any type of variable into the analysis, b) Dynamic, as they can easily accommodate any information as they become available to adjust the model accordingly. A recent study shows that around 21 million Americans are diagnosed with asthma, which represents a large section of the insurance and pharmaceutical patient population. Asthma patients are classified into two groups: Controlled vs. Uncontrolled asthma; however, during the year 2000, the national surveys show that there were nine million physician office visits and more than 900,000 hospitalization cases in addition to the emergency room visits. Based on the American Lung Association reports, in 2007, almost 2.5 million people over the age of 65 had asthma, and more than 1 million had an asthma attack or episode.*(American Lung Association 2007)* We examine data from Medical Expenditure Panel Survey (MPES) and California Health Interview Survey. It was proven that the use of predictive modeling did show benefit by pointing to optimal treatment for patients and can be used to develop better investments for health plans.

INTRODUCTION

Asthma is a chronic inflammatory disorder of the airways, designated as ICD9-CM-493 in the (ICD-9 classification system. Asthma is a chronic disease that cannot be cured. Asthma is ranked as one the most common diseases within the United States. It is usually treated by two types of medications: long term and quick relief medication. Missing work, missed school days are examples for the indirect cost while treatments and hospitalization represent the direct cost. Asthma patients use a significant portion of insurance companies, state insurance funds and hospital resources. Asthma hospitalization rates are growing, especially for children. Since Asthma patient are classified as high cost patients, as the Urban Institute's health costs survey estimates that an asthma patient can spend as much as $4,900 many research studies were directed at hospitalizations to decrease the healthcare bill. (*Health Costs Survey 2005*)

Medications are the prime cost components for the asthma patient; for example, for inpatients, asthma represents 30% of the overall cost. Asthma guidelines generally use an Expert panel, and have been developing over the years to the current state as consensus panel guidelines. Since guidelines rely largely on observational studies, the definitions are shaped by consensus; while it seems reasonable that there should be a relationship between observations and consensus, this does not necessarily occur. On the other hand, predictive modeling provides a great opportunity for better healthcare planning, management and clinical practice guidelines. Identifying the patients who will need medical attention in the future enhances health care management guidelines, which exist to define standard treatments for patient and to reduce the utilization of resources.

Generally speaking, asthma affects children and African Americans within the general population and affects females more than males within the adult population. Data collected over the last few years show that asthma is ranked as one of the top illnesses that affect the American people. Almost 15 million people are diagnosed with asthma, The National Heart, Lung and Blood Institute states that costs for Asthma are approximately $14.0 billion with $9.4billion in direct costs, $2.7billion for morbidity, and $1.9billion for mortality.

Adults or children who are diagnosed with asthma miss work, school days, and also sleep poorly at night, and in many cases, asthma can affect their mobility. On the other hand, several things can affect the patient's general health condition such as nutrition, smoking, and exercise. Uncontrolled asthma can lead to several complications such as school absenteeism, breathing difficulty, hospitalization and death, in some cases.

The main focus of service providers, payers and decision-makers is to minimize the number of emergency room visits and the high cost of hospitalizations. Several intervention approaches are used to ensure that goal. However,

the large number of patients highlights the need for efficient resource management and the need to predict the future health requirements to avoid a resources shortage, as the main challenge is to categorize patients who will need treatment. To achieve this goal, decision-makers want to investigate several issues, such as the number of patients with a controlled condition vs. uncontrolled, geographical patient distribution, interventions and so on.

Predictive modeling is a powerful tool to enhance the decision making process; in the healthcare domain, it is a relatively new concept. Predictive modeling calculates the probability of future adverse events for any patient based on past clinical information compared to the population. The model we define is created using several factors, or "predictors," which are expected to affect the future of this patient status, such as age, gender, and medical history. Predictive modeling tools are used for several applications such as:

1- Predicting future costs and the design of benefits packages
    a. deductibles
    b. co-pays
2- Patient profiling
    a. Identification of high-risk patients
    b. Disease management and case management
    c. cost savings
3- Provider profiling/Plan profiling
4- Risk adjustment

For the last few years, predictive modeling accuracy has increased and proven to improve healthcare management. It can be used to develop better investments for health plans and enhance resources utilization. The strength areas of predictive modeling are:

a) The ability to incorporate any type of variable into the analysis
b) Dynamic, as a model can easily accommodate any information as it becomes available to adjust the model accordingly.

USING PREDICTIVE MODELING

Predictive Modeling is one of the two main branches of data modeling; it is a process to create a statistical model of future behavior. Regression and Classification demonstrate the main idea of predictive modeling *(Cerrito 2006)*

- If the Response variable is numerical, predictive modeling is called regression.
- If the Response variable is nominal, predictive modeling is called Classification.

Developing predictive models includes data preparation, prediction, and the analysis of results. The model tuning process is a time consuming processes. The main challenges that the data analyst encounters are

- Selecting the predictive method
- Selecting the independent variables, or *predictors*

Regression models are the backbone of predictive modeling, Regression is the process of establishing a mathematical model as a function to represent the relation between the different variables (predictors). There is a wide-range of models that can be applied while performing predictive analytics. Some of them are:

- Linear regression model
- Logistic regression
- Multinomial logistic regression

Artificial Neural networks (ANN) provide an information processing model for computing that involves developing mathematical structures with learning. ANNs are used to provide projections given new situations of interest and to answer "what if" question. The following figure shows the ANN diagram.

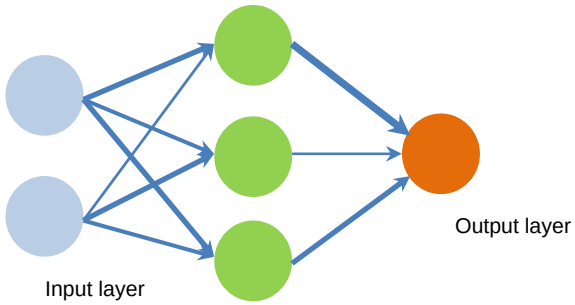

Figure 1. Artificial Neural networks

Decision Tree is a predictive modeling tool that does prediction by classification approach. The key element for building a decision tree is finding the best rule that can spilt the data record in the best manner. At each level, the best splitting rule is applied until you cannot find better ways to split the input records.

SAS® SEMMA METHODOLOGY

The SEMMA methodology was introduced by SAS® technology; SEMMA represents the lifecycle of the core process of data mining. It also represents the complete functionally of SAS Enterprise Miner, and is acronym for sample, explore, modify, model, and assess; SEMMA can be described briefly as follows:*(Sumathi, 2006)*

- *Sample*: Extracting a portion of a large data set big enough to contain the significant information, but still manageable. In addition, it supports a data partition into Training, Validation, and Testing subsets.
- *Explore*: To enable you to become more familiar with your data and to detect any pattern or any anomalies. This process can be done visually or using techniques such as clustering, path analysis and text analysis.
- *Modify*: Data may need to be modified for a number of reasons; since data mining is a dynamic process, then the objective can change and the data will be modified. The other reason to modify data is to utilize data discoveries.
- *Model:* it the process of using current data to investigate the relationship between the parameters to create a model capable of future behavior. Several techniques can be used such as neural network, tree based and regression models.
- *Assess*: it is very important to be able to test the usefulness and reliability of a model, which can be done easily by testing model against known data.

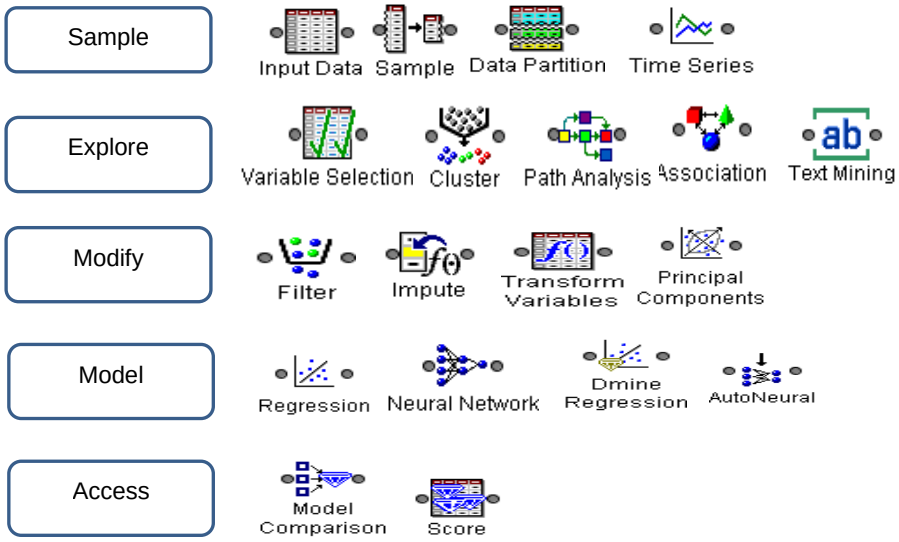

Figure 2. SAS® SEMMA METHODOLOGY

DATA SUMMAR

To illustrate the asthma hospitalizations risk problem, the Medical Expenditure Panel Survey (MEPS) data sets were used to find general conclusions about asthma in the United States and the California Health Interview Survey (CHIS) data sets were used to investigate predictive modeling. As shown by the following table (Table 1) 13.6% of the population has been diagnosed with asthma at least once based on the CHIS 2009 data sets. (http://www.chis.ucla.edu/)

For this study we focused on the CHIPS dataset for the year 2009. It surveyed 47,614 adults, 3,379 adolescents and 8,945 children. The survey includes robust samples of various races. We tested several subsets of parameters or factors to be tested versus the hospitalization risk or need for emergency room visits or the frequency of attacks. In this study, we included many parameters as follows:

- Demographic Information: Age, Gender, Race, Marital Status.
- Health Behaviors: physical activities, Fast food, Alcohol consumption.
- Health Conditions other than Asthma.
- Health Insurance.
- Poverty level.
- Emergency Preparedness Module: Medication
- Mental or emotional condition.

Table1. Patient diagnosed with asthma summary

	Male		**Female**		**All**	
Ever diagnosed with asthma	No. of Obs.	%	No. of Obs.	%	No. of Obs.	%
Has asthma	2,401,000	13.4	2,543,000	13.9	4,944,000	13.6
Does not have asthma	15,574,000	86.6	15,747,000	86.1	31,321,000	86.4
TOTAL	17,975,000	100	18,290,000	100	36,265,000	100

The following figure (Figure 1) shows that the number of females diagnosed with asthma is higher than for males. Also, the number of females who no longer have asthma is smaller compared to the number for males.

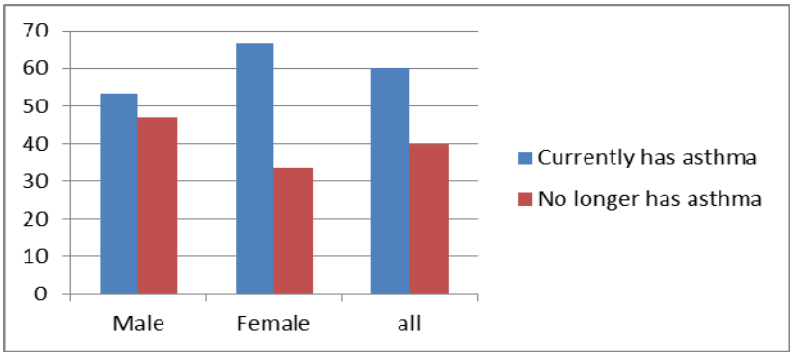

Figure 3. Patient diagnosed with asthma at least once

Another aspect to check is marital status, based on the CHIS classification, which uses four main categories as shown in the following table (Table 2).

Table 2: Marital status summary

	Male		**Female**		**All**	
Marital status	No. of Obs.	%	No. of Obs.	%	No. of Obs.	%
Married	7,461,000	56.6	7,316,000	53.4	14,777,000	55
Live with partner	1,002,000	7.6	976,000	7.1	1,978,000	7.4
Separated / Divorced / Widowed	1,225,000	9.3	2,641,000	19.3	3,866,000	14.4
Single, never married	3,490,000	26.5	2,763,000	20.2	6,253,000	23.3
TOTAL	13,179,000	100	13,695,000	100	26,874,000	100

The insurance coverage is a very important element for the patient to maintain an acceptable level of medical treatment, in particular, for out-patients. The following table (Table 3) represents summary for patients who had no insurance at all for the last 12 months, either partial or complete coverage.

Table 3: Medical insurance coverage summary

During the last 12 months	No insurance at all	Insurance for a part	Had insurance entire time
Currently have asthma	%	%	%
Currently has asthma	48.8	56.7	60.6
No longer has asthma	51.2	43.3	39.4

It is known that the asthma patient can experience work or other activities limitations; the following table (Table 4) illustrates the scope of the problem. For example, 58.8% of the patients who had 1-2 days of work limitation had it caused by asthma.

Table 4: Work/other activates limitation caused by asthma

During the last month	None	1 - 2 days	3 - 5 days	6 - 12 days	13 - 29 days	30 days
Caused by asthma	59.1%	58.8%	61.3%	65.9%	64.3%	67.1%

Data show that children and senior patients are more affected by asthma as shown in the following table (Table 5).

Table 5: Age categories

During the last month	Child (0-11)	Adolescent (12-17)	Adult (18-64)	Senior (65+)
Diagnosed asthma	65.4%	57.8%	58.2%	66.4%

USING SAS® PREDICTIVE MODELING

The following figure (Figure 4) shows the data diagram for predictive modeling using SAS® Enterprise Miner. As a data preprocessing step, data were filtered to include only patients diagnosed with Asthma using the ICD-9 code of 493. Data were partitioned using the data partition node with percentages 40% training, 30% testing, and 30% validation. *(Sarma, 2006)*

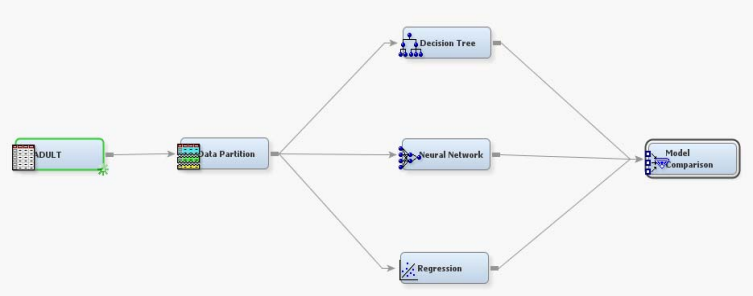

Figure 4. Predictive Modeling Diagram Note: use version 5 or 6, not version 4.3 for this diagram

One of the major problems of using predictive modeling is that for some applications, certain models will not work, yet SAS software provides an easy way to use several models such as neural network, regression and decision. In addition, it provides an assessment node to compare each model result and to chose the best one.

For the first round, we used all of the 473 response parameters as input and the occurrence of hospitalization as the outcome variable. It showed that several factors were related:

- ER visits within the last year
- Asthma episode/attacks in the past 12 months
- How often had asthma symptoms
- Others

It is valid that those parameters have influence on the occurrence of the hospitalization event; however, this study's objective is to go beyond that and to investigate other factors affecting the hospitalization event. Within the decision

tree model, a competing splitting feature provides an easy way to explore other splitting factors. For example, the following table (Table 6) shows the competing splitting vs. the ER visits with the past year.

Table 6: Competing splitting

Variable	Logworth	Groups	Label
ER	*24.694*	*2*	*ER VISIT WITHIN THE PAST YEAR*
AB1	13.429	2	GENERAL HEALTH CONDITION
AB42	12.731	2	WORKDAYS MISSED DUE TO ASTHMA IN PAS
AD52	10.549	2	HAS DIFFICULTY DRESSING, BATHING, GETTI
AB41	9.134	2	ASTHMA EPISODE/ATTACK IN PAST 12 MOS

Logworth is a split decision statistic that measures the effectiveness of a particular split decision at differentiating values of the target variable.

For the next round of the run, the "*expected*" factors were removed to give more opportunity for the competing splitting. The second round showed parameters such as

- Taking daily medication to control asthma
- General health conditions
- Type 1 or Type 2 diabetes
- Psychological Distress
- Poverty level

To achieve a better understanding of the problem, we used a backtracking technique. For instance, tracking daily medication to control asthma is very significant to avoid future hospitalizations. Rx-Coverage, disability, language barriers and some mental problem can prevent asthma patients from taking the medication regularly.

The following figures show sample output and results for predictive modeling models: Decision tree, Regression and neural networks.

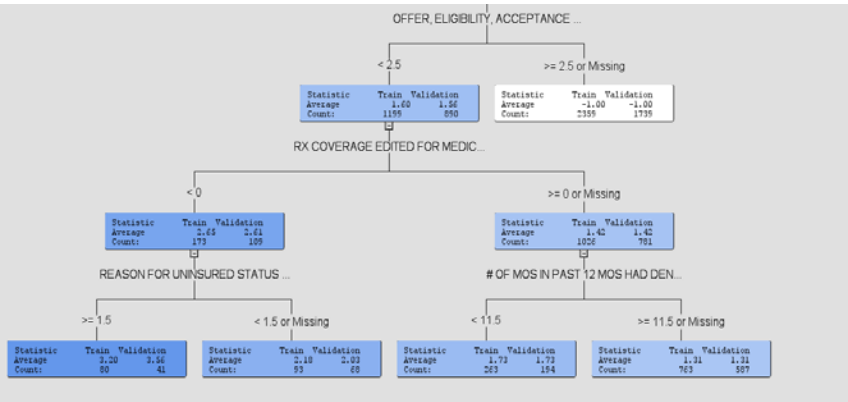

Figure 5. Sample output of decision tree model

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
AI15	_DFT_	Total Degre...	20419	.	.
AI15	_DFE_	Degrees of ...	19032	.	.
AI15	_DFM_	Model Degr...	1387	.	.
AI15	_NW_	Number of ...	1387	.	.
AI15	_AIC_	Akaike's Inf...	-32057	.	.
AI15	_SBC_	Schwarz's ...	-21066.1	.	.
AI15	_ASE_	Average Sq...	0.181625	0.229409	0.217403
AI15	_MAX_	Maximum A...	6.633353	7.143714	6.484293
AI15	_DIV_	Divisor for A...	20419	15314	15315
AI15	_NOBS_	Sum of Fre...	20419	15314	15315
AI15	_RASE_	Root Avera...	0.426174	0.478966	0.466265
AI15	_SSE_	Sum of Squ...	3708.594	3513.166	3329.521
AI15	_SUMW_	Sum of Cas...	20419	15314	15315
AI15	_FPE_	Final Predic...	0.208097	.	.
AI15	_MSE_	Mean Squa...	0.194861	0.229409	0.217403
AI15	_RFPE_	Root Final ...	0.456177	.	.
AI15	_RMSE_	Root Mean ...	0.441431	0.478966	0.466265
AI15	_AVERR_	Average Err...	0.181625	0.229409	0.217403
AI15	_ERR_	Error Functi...	3708.594	3513.166	3329.521
AI15	_MISC_	Misclassific...	.	.	.
AI15	_WRONG_	Number of ...	.	.	.

Figure 6. Fit statistics for regression model

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
AI15	_DFT_	Total Degre...	20419	.	.
AI15	_DFE_	Degrees of ...	19032	.	.
AI15	_DFM_	Model Degr...	1387	.	.
AI15	_NW_	Number of ...	1387	.	.
AI15	_AIC_	Akaike's Inf...	-32057	.	.
AI15	_SBC_	Schwarz's ...	-21066.1	.	.
AI15	_ASE_	Average Sq...	0.181625	0.229409	0.217403
AI15	_MAX_	Maximum A...	6.633353	7.143714	6.484293
AI15	_DIV_	Divisor for A...	20419	15314	15315
AI15	_NOBS_	Sum of Fre...	20419	15314	15315
AI15	_RASE_	Root Avera...	0.426174	0.478966	0.466265
AI15	_SSE_	Sum of Squ...	3708.594	3513.166	3329.521
AI15	_SUMW_	Sum of Cas...	20419	15314	15315
AI15	_FPE_	Final Predic...	0.208097	.	.
AI15	_MSE_	Mean Squa...	0.194861	0.229409	0.217403
AI15	_RFPE_	Root Final ...	0.456177	.	.
AI15	_RMSE_	Root Mean ...	0.441431	0.478966	0.466265
AI15	_AVERR_	Average Err...	0.181625	0.229409	0.217403
AI15	_ERR_	Error Functi...	3708.594	3513.166	3329.521
AI15	_MISC_	Misclassific...	.	.	.
AI15	_WRONG_	Number of ...	.	.	.

Figure 7 Fit statistics for neural network model

The above figures (Figure 6 & Figure 7) show fit statstics for the neural network model and the regression model, selected statstics can be used as the modeling selection criteria statistics such as: average profit and misclassification rate. While the average error is the modeling selection criteria statstics for chosing the best neural network mode.

If the modeling t statistics are totlay changed between the partitioned data sets then it will be an indication of overfitting to the regression model.

Another important feature of a SAS® Enterprise Miner solution offers is the assessment node, where it compares ASE (Average Square Errors), Schwarz Bayesian criterion and the misclassification rate. In many cases, the neural network model will have a slightly smaller ASE and misclassification rate. The main concern with neural network is that they are hard to explained.

CONCLUSION

SAS® Enterprise Miner provides a powerful environment for predicative modeling by supporting several models and the ability to compare between them. The use of predicative modeling for solving healthcare problems is not new, yet this study illustrates a new application, risk of future hospitalization for asthma patients. The results show that predictive modeling can enhance the decision making process and future planning for hospitals' resource utilization. Also, it was clear that by using predictive modeling results in terms of which parameters influence future hospitalizations, it is possible to enhance interventions, programs and alternatives to avoid future hospitalizations. Results show that the general health conditions, psychological distress and poverty level affect future hospitalization risk. In addition, the Rx coverage and patient's disability influence taking medication regularly and can increase the future hospitalization risk.

REFERENCES

1. American Lung Association. Epidemiology & Statistics Unit, Research and Program Services. Trends in Asthma Morbidity and Mortality, November 2007.
2. Health Costs Survey 2005.
3. Patricia B. Cerrito: Introduction to data mining using SAS Enterprise Miner, SAS Institute, 2006.
4. S. Sumathi, S. N. Sivanandam: :Introduction to data mining and its applications, Springer 2006.
5. Kattamuri S. Sarma: Predictive Modeling With SAS Enterprise Miner: Practical Solutions for Business Applications, SAS Institute , 2006.
6. National Asthma Education and Prevention Program Expert Panel Report 3: *Guidelines for the Diagnosis and Management of Asthma.* Rockville, MD. National Heart, Lung, and Blood Institute, US Department of Health and Human Services; 2007. NIH publication 08-4051.
7. Woodfield, Terry: "Data Conversion Pitfalls", Proceedings of Computer Science and Statistics: 26th Symposium on the Interface, 1994, pp. 362-371.
8. MC Weinstein, EL Toy, EA Sandberg, PJ Neumann, JS: Modeling for Health Care and Other Policy Decisions: Uses, Roles, and Validity, Value in Health, Blackwell Synergy, 2001.

ACKNOWLEDGMENT

The author would like to thanks Prof. Patricia Cerrito of the University of Louisville for her helpful suggestions and advices.

CONTACT INFORMATION

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

Yehia H. Khalil
University of Louisville, Louisville KY
Email: Yehia.khalil@ieee.org