PharmaSUG 2014 - Paper PO17

Healthcare Data Manipulation and Analytics Using SAS

Lumin Shen, University of Pennsylvania

Jane Lu, AstraZeneca Pharmaceuticals Inc

ABSTRACT

Increasing application of information technology in the healthcare delivery system helps healthcare industry gain valuable knowledge from data, and use this insight to recommend action or guide decision making, and improve quality of patient care and practice of medicine. There is more data available than ever before. How can it truly benefit patients, payers and healthcare providers? Analytics can help medical researchers exploit healthcare data to discover knowledge lying implicitly in individual patients' health records, physicians identify effective treatments and best practices, patients receive better and more affordable healthcare services, and healthcare insurance companies detect fraud and abuse. This article explores analytics applications in healthcare industry. It illustrates the process of data integration and exploration, and building of predictive models to find previously unknown patterns and trends. It also presents analytics applications in major healthcare areas.

K E Y W O R D S

■ Healthcare analytics
■ ETL
■ Healthcare applications
■ Statistics methodology and techniques
■ Predictive modeling

INTRODUCTION

Healthcare organizations generally adopt information technology to save time, reduce costs or errors as well as improve efficiency and quality by automating patient care transactions, such as scheduling appointments, filling or refilling prescriptions, lab testing and billing. These benefits are only icing on the cake. The widespread adoption of EMR systems, (and in turn, a national EHR database) makes large quantities of healthcare data available. Healthcare analytics is becoming increasingly popular, and analytics applications can greatly benefit all parties involved in the healthcare industry.

This article explores analytics applications in healthcare. It illustrates the process of data integration and exploration, and building predictive models using vast database to find previously unknown patterns and trends. It also presents analytics applications within healthcare in major areas such as

- Management of healthcare such as physician's offices, hospital beds, retail pharmacies, nursing homes and medical equipment rentals
- Direct marketing for marketers, market segmentation for retailers
- Evaluation of treatment effectiveness
- Identification of risk factors associated with diabetes and coronary heart diseases
- Customer relationship management
- Detection of insurance fraud and abuse

These analyses have become increasingly essential for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from analytics can influence cost, revenue, and operating efficiency while maintaining a high level of care.

Healthcare analytics involves business understanding, data understanding and preparation, modeling, evaluation, and deployment.

- What treatment regimen yields the best outcomes for patients with the genetic profile?
- What insidious drug interactions are we likely to see in a patient with these risk factors?

- How well does this combination of therapies help patients undergoing the procedure?
- What protocol produces the best rehabilitation results for the target population?

To answer questions like these, researchers need business intelligence (BI). BI has been used for years to run the business better. Business understanding is critical because it identifies business objectives, and thus, the success criteria of analytics projects.

Now, with new sources of digital data about patients and their clinical experiences, it is important to explore, prepare and visualize the data before modeling.

DATA CHARACTERISTICS

Healthcare information systems typically reflect a process- and patient-oriented view of the business. The raw inputs of healthcare data often exist in different settings and systems, such as administration, clinics, laboratories, financial, operational, research and so on. The architecture uses a host of independent systems on different platforms, which share information in a limited way, if at all. As growth, mergers and acquisitions reshape information networks, it's common to see multiple, incompatible platforms even within a single functional area.

Like any other industries, healthcare industry has its own jargons. There are hundreds of thousands of acronyms, codes, and special terminology. The Current Procedural Terminology (CPT) code set is a medical code set maintained by the American Medical Association through the CPT Editorial Panel[1] The CPT code set (copyright protected by the AMA) describes medical, surgical, and diagnostic services and is designed to communicate uniform information about medical services and procedures among physicians, coders, patients, accreditation organizations, and payers for administrative, financial, and analytical purposes. With the changes in healthcare practice, new codes are developed for new services, current codes may be revised, and old, unused codes may become obsolete and discarded. Thousands of codes are being used, and updated annually.
Examples of CPT Codes:

- 90658 indicates a flu shot
- 55873 indicates Cryosurgical ablation of the prostate
- 90716 refers to chicken pox vaccine (varicella)
- 12002 stands for a one-inch cut on a patient's arm

Medicare uses a slightly different coding system known as HCPCS - Healthcare Common Procedure Coding System and is monitored by CMS, the Centers for Medicare and Medicaid Services. HCPCS Codes are numbers assigned to medical services provided to a Medicare patient. There are two sets of HCPCS codes. The first set, Level I HCPCS codes, are identical to CPT codes. Level II HCPCS codes are alphanumeric medical diagnostic codes, primarily for non-physician services such as ambulance services and prosthetic devices.

The International Classifications of Diseases (ICD) is the standard diagnostic tool for epidemiology, health management and clinical purposes. It is developed, monitored and copyrighted by the World Health Organization (WHO). It is used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and health records. The Clinical Modification of the codes (ICD-9-CM) provides additional fourth and fifth digits to the rubric to allow more detailed reporting. Each diagnosis on human beings has a code, a numbered designation. For example, if diagnosed with GERD (acid reflux), it will be given the code 530.81.

With few exceptions, the paperwork from a doctor's office will contain both CPT codes to describe the service that is rendered for billing purposes, and ICD-9-CM codes to describe why that service is provided. As we move more and more into electronic medical records, these codes will be used more frequently by physicians and other medical professionals. Many of these coding systems are centrally organized and well documented. It is important to know the meaning of the codes and how to interpret the study results correctly.

HIPAA (Health Insurance Portability and Accountability Act of 1996) recognizes the need for healthcare organizations to share patient data, but requires to maintain privacy and security of individually identifiable health information. HIPAA Privacy Standard considers information to have been de-identified by removing 18 explicit elements of data such as name, address, phone number, SSN, etc., before doing any analysis. De-identification also requires to create a unique anonymous identifier for each patient. Such de-identified health information can then be used or disclosed without restriction.

Other challenges in healthcare data are the large volume, complexity and heterogeneity of medical data and their poor mathematical characterization and non-canonical form. These include missing, corrupted, inconsistent, or non-standardized data. In particular, the lack of a standard clinical vocabulary is a serious hindrance to data analysis.

DATA EXTRACTION/TRANSFORM/LOAD

SAS is a very powerful analytical tool, but not all healthcare data comes as a SAS dataset. Being able to get data into SAS is an integral part of the work. Generally, data warehousing environment and analytical environment are kept on separate platforms to maximize resources for both, which has made communication between SAS and database challenging.

ETL(export, transform and load), describes the process of “exporting” data from a source, performing some modification of the data in “transforming”; and finally “loading” the data into another system.

One of the data transfer methods is to export and import the data through intermediate delimited files such as CSV files. It is simple and convenient for small data, but doing so for large data had many obvious drawbacks, including efficiency, real-time data cuts and data integrity.

Getting SAS to interact with other databases is often vital to the success of a project, but it won’t be easy when SAS and the databases being used are on different platforms. Oracle, MySQL, MS SQL Server, Netezza, Teradata and DB2 are generally used relational databases. A relational database is tables that are related to each other via fields that are shared. The following methods are commonly used for SAS to databases

(1). Pass Through Facility uses Proc SQL and requires all the code that will be passed to the databases to be written in their native SQL language (e.g., Oracle's PL/SQL, Microsoft's T-SQL). This enables the user to tap into the relationships, indexes, and other settings supplied by database and have the resulting dataset returned in SAS format. Oracle database as an example for processing:

```
PROC SQL;
  CONNECT to ORACLE ( USER = cschacherer PASSWORD=tiger PATH = billing);
    CREATE TABLE july_charge_summary AS
       SELECT * FROM CONNECTION TO ORACLE
       (SELECT TO_NUMBER(patient_number), SUM(charges) as charges
         FROM charge_detail
           WHERE bill_date BETWEEN TO_DATE('07/01/2010','MM/DD/YYYY') AND
              TO_DATE('07/31/2010','MM/DD/YYYY')
                GROUP BY TO_NUMBER(patient_number));
    DISCONNECT FROM ORACLE;
QUIT;
```

(2). ODBC CONNECTION (Open Database Connectivity) allows a programmer to access data from database management system with a non-native application. ODBC connection can be used with JMP, MS-ACCESS, SAS, SQL-SERVER, Verity Teleforms and in web applications.

Microsoft has installed an ODBC data source administrator, a standard odbc wizard, which makes connecting straightforward. This is where you create the odbc connection and assign a database system name (DSN) to it, which is how you tell SAS which odbc connection to use. LIBNAME statement is then used for data connection:

```
libname save odbc dsn=oracle uid=scott pwd=tiger;
```

```
LIBNAME SQL ODBC DSN=’sql server’ user=sasjlb pw=pwd;
```

Where 'sql server' is the name of the Data Source configured in the ODBC Administrator.

SQL Server database tables are organized in schemas, which are equivalent to database users or owners. In order to see particular tables in a defined library, you may need to add the SCHEMA= option to the LIBNAME statement. If no schema is specified, SAS will look in the current userid’s schema by default. For example:

```
LIBNAME SQL ODBC DSN=sqlsrv user=sasjlb pw=pwd schema=dbo;
```

With OLEDB connection, it is not necessary to configure the data provider. The LIBNAME statement would look similar:

```
LIBNAME sqlsrv oledb init_string="Provider=SQLOLEDB.1;Password=pwd;
Persist Security Info=True;User ID=user;Data Source=sqlserv";
```

(3). SAS/ACCESS: SAS Version 9 has SAS/ACCESS interfaces specifically for Oracle, SQL Server, MySQL, Netezza and Teradata. They allow users for reading and writing data with database. The connection string is supplied using LIBNAME statement:

```
LIBNAME hds NETEZZA dsn=RD_DEV uid=svc-sas_nz pwd=$AAA2012 connection=shared;
```

DATA CLEANUP / DESCRIPTION

Proc SQL, ODBC and SAS/ACCESS enable users to pull data across networks quickly and efficiently. End users can have almost real time access to the collected data. It is especially useful in checking the accuracy of the data from patient health studies and clinical trials.
Data quality and error check would be done first as the data is collected and put into SAS datasets.

As mentioned before, HIPAA requires to maintain the privacy and security of individually identifiable health information. Acronym is used to describe Protected Health Information as defined in the HIPAA Privacy Rule 164.514 regulations pertaining to a patient, including but not limited to names, birthdates, account numbers or geographical metadata. The detailed description for each data element will be deleted or masked to create de-identified data.

To ensure patient privacy and minimize potential risks, all de-identified patient data to be shared will be verified using a SAS macro with VERIFY() function to enhance de-identification. The details of the coding is available in Appendix.

```
%check_masking  (lib=xdata, outdsn=varchk);
```

Common data issues are:

- Missing: ~10% of all records have blank, missing or default values in key data fields: DOB, gender, race.
- Duplicate record: a patient has two or more assigned MRNs, average duplication rate is 8%-12%.
- Overlap records: a patient has different MRNs in separate facilities that are linked in one EMRI, one MRN contains information on two separate individuals.
- Confidentiality: potential confidentiality breached, cost of litigation.

It's a challenge to reconcile and clean all this incompatible data, much less to gain useful information from it. Queries on database are used to verify de-identification, to remove duplicates, to ensure integrity, to identify data anomalies and monitor data values and attributes such as validity, completeness, consistency, reliability.

A variety of SAS functions, procedures and techniques increase the cleanliness of data, real-time quality check, and identify root causes of problems and integrity issues. PROC FREQ and PROC MEANS/PROC UNIVARIATE are among the most heavily used PROCs to detect and find unusual data and outliers.

PROC FREQ is used to produce frequency distributions and perform inferential statistical tests on categorical data.

PROC MEANS provides descriptive statistics such as Mean, Median, Standard Deviation, Min./Maxfor continuous, numeric variables. It is also a valuable data cleaning tool—as it can reveal out-of-range values (e.g., revealing variables like "Cost", and "Billed Amount" that may have too little or too much variance based on historical data, etc.).

PROC UNIVARIATE provides a more comprehensive set of descriptive statistics, such as Quartile break (which show population breaks into 0 – 49, 50 – 60, 61 – 70, 70-89 and 90 & above in roughly equal numbers), stem-leaf plot and the corresponding smoothed histogram.

```
PROC UNIVARIATE DATA=patients;
VAR age;
HISTOGRAM age /normal ;
RUN;
```

PROC SORT with OPTIONS of NODUPKEY or RETAIN in conjunction with the DATA STEP are important techniques for efficiently removing duplicates and cleaning dirty data.

DATA VISUALIZATION

Descriptive statistics and visualization can help understand a data set, especially a large one, and detect hidden patterns in data, especially complicated data containing complex and non-linear interactions.

SAS/SG procedures—SGPLOT, SGPANEL, and SGSCATTER—provide new tools for viewing and reporting data. Various plots are to help understand the counts of visits and patients trends within therapeutic areas or specified population groups, to construct laboratory panel using treatment regimen and visit numbers as the classification variables, a matrix of liver function tests (LFTs) for at-risk patients and aggregate data into on-the-fly classification variables using user-defined formats. In addition, PROC KDE allows the investigator to overlay smoothed histograms to examine data. It is available for interval data only. In Enterprise Miner, it is possible to use histograms to examine the data.

The SG procedures are an extension of the ODS Graphics framework, providing access to the Graph Template Language (GTL) in the familiar syntax of the SAS/GRAPH procedure. Starting with SAS 9.2, SAS provides user with the ability of GTL to generate professional looking graphic output and store the template for reuse with other data. GTL is the underlying language for the default templates that are provided by SAS for procedures that use ODS Statistical Graphics and is very powerful.

With large datasets, concise visual presentation becomes an important part of exploring data. They are usually performed prior to modeling.

ANALYTICS APPLICATIONS

Even when healthcare data is available in digital form, it is usually for billing purposes rather than for direct analysis. Some of this data is episodic and patient-focused, and is cumulative over time, entities and populations.

Healthcare data are built with information used to record patient admissions and discharges, bill patients and insurers, order laboratory and radiological tests and dispense medications.
SAS provides most data manipulation and analysis procedures for analytical processing and statistical methods, such as regression analysis and cluster analysis, discriminant analysis and using neural networks, decision trees, link analysis and association analysis.

1).TREATMENT EFFECTIVENESS

Clinicians are often interested in evaluating the effectiveness of medical treatments. The outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatment works best and is most cost-effective.

By comparing and contrasting disease symptoms (e.g., tumor type, AJCC staging, time to progression), patient conditions (e.g., gender, age, comorbidities) and courses of treatments (e.g., agents including orals, regimen, line of therapy, dosing schedule, supportive care), PROC MIXED can explore which courses of action prove more effective. It also can develop clinical profiles to give physicians information about their practice patterns and to compare these with those of other physicians and peer-reviewed industry standards.

As healthcare companies continue to invest in new therapies and treatments to improve oncology care, we've made this a priority. By covering the full patient treatment history of anti-cancer therapies, our oncology analytics and insights help clients identify valuable market opportunities, evaluate competitive behavior and optimize their portfolios.

Physicians want to evaluate clinical effectiveness of Valrubicin via Radical Cystectomy for patients with BCG refractory non-muscle invasive bladder cancer, by comparing the survival status of 3, 6, 12 months post therapy. Survival analysis procedure LIFETEST will do the analysis based on the patient longitudinal data:

```
proc lifetest data=onctte timelist=(3,6,12);
  time month * censor (0) ;
```

```
  strata treatment;
run;
```

To evaluate the impact on the overall survival by the following covariates:

- Treatment
- Tumor types (CIS alone vs. CIS plus papillary)
- Duration and extent of prior BCG
- Time from first bladder cancer diagnosis to initiation of therapy

PROC PHREG with Cox-hazard regression model will be used to generate the hazard rations of each covariate with 95% confidence intervals, p-value for each covariate and number of observations used in the models will be presented.

2).CLUSTER ANALYSIS

Analytics applications can develop to identify and understand high-risk patients, and high-cost patients, in order to design appropriate interventions, and reduce the number of hospital admissions and claims. With clustering, patient populations are stratified by demographic characteristics and medical conditions, in such a way that similar patients belonging to the same cluster, dissimilar patients to different clusters .

Clustering involves the grouping of observations based upon the distance between observations. Distance is defined using different criteria available in PROC CLUSTER. The clusters are built on a hierarchical tree structure. The closest observations are grouped together initially followed by the next closest match. Clustering involves the following process:

- Group observations or group variables
- Choice of type-hierarchical or non-hierarchical
- Number of clusters
- Cluster identity
- Validation

The number of clusters defined range from two to many, make a judgment as to which is optimal. Once the hierarchy is established and the number of clusters is chosen, the final clusters can be labeled a meaningful identity. The other method of clustering is based upon the selection of random seed values as the centers of spheres containing all observations closest to that center (PROC FASTCLUS). Then the centers are re-defined based upon the outcomes. PROC FASTCLUS is used to perform clustering. In addition, a multidimensional scaling (PROC MDS) is performed to examine the separation between clusters. The scaling is based upon distance within and between clusters. Several distance criteria can be used for comparison purposes.

Similar to clustering, segmentation is a technique used in marketing and sales management to create groups that have similar characteristics. It can also give a measure of how ‘different’ one group is from another. Segmentation can be used to answer questions such as:

What are the attributes of a typical Brand A prescriber compared to other physicians?

Which group of physicians are good prospects for surgery operations?

As large databases are used more and more, pattern recognition gains significant importance for strategic marketing decisions, market segmentation, and sales promotion effectiveness.

Segmentation is used to identify the prescribers who have changed their prescribing behaviors: Brand Switching, Brand Loyalty and to understand the influence of factors that caused these pattern changes. The method enables sales force to pinpoint prescribers who are switching from one medication to another. A sales person can target doctors who have switched from the drug they are selling and to devise a specific message to counter that switching behavior.

3).TREND ANALYSIS

Marketing-research analysts, physicians, financial analysts, and others, study time series to understand general medical market trends and take action. For example, physicians would like to know the number of office visit and patient for the last 5 years relating to CPT code, broken out by year or even month for each year. Analytic approach can outline the longitudinal data into various trends: Increasing, Decreasing, No Pattern, and even broken out by Medicare, Medicaid, HMO or private insurers, and payment of cash.

In analyzing reimbursement trends for a diagnostic code, PROC REG and AUTOREG can estimate and predict the target variable, such as predicting the length of hospital stay for a particular condition and procedure or the amount of resource utilization.

Variable Selection is a technique used to reduce the number of possible variables that are used in modeling step. This is commonly used to avoid over-fitting the data to the model, where too much information is used in building the model so that the model is only useful for describing the sample data and not for the general population. It is also more generally true that a model with fewer parameters that adequately predicts the result is more desirable than a more complex model. This is because there is often a cost associated with collecting the data needed to feed the model and it is generally easier to explain. A community hospital team used SAS to track the long-term care and rehabilitation of spinal surgery patients, to demonstrate to Medicare that although upfront costs were high, cumulative costs were lower than other hospitals and patients returned to productive lives in the community.

Forecasting is an important topic to keep financial performance on track at a time when revenues are less predictable. High performance forecasting can be applied to healthcare data. SAS/STAT software facilitates data processing and forecasting procedures to provide useful information that can help improve patient care and lower costs.

Variables used include medication start date, medication quantity, patient insurance payment, and total cost of medication. SAS High Performance Forecasting (PROC HPFDIAGNOSE & PROC HPFENGINE) are used to investigate trends in medication prescription, usage, and cost. The medications were collected at a series of irregular intervals. PROC HPFENGINE automatically selects the IDM (Intermittent demand model) if most of the observations in the data are missing. As a result, ARIMAX model, UCM (unobserved component model) and ESM (exponential smoothing model) will be fitted to the longitudinally prescribing medication datasets to forecast the trends in prescription practices such as total cost, private insurance payment, quantity of prescribed medication and the number of prescriptions.

For predictive modeling, the analytical techniques include multiple discriminant analysis and logistic regression analysis.

4).CLASSIFICATIONAND PREDICTIVE MODELING

Classification analysis reveals hidden patterns in masses of data, without having a predetermined idea or hypothesis about what the pattern may be, for example, predicting the high-risk individuals such as diabetes, or the response to changes in treatment protocols for ventilator patients. Applying sophisticated models and algorithms in the process uncover very important patterns and best practices.

PROC DISCRIM uses the following seven variables of particular interest: gender, age, body mass index (BMI), waist hip ratio (WHR), smoking status, number of times a patient exercises per week, and the target variables - onset of diabetes, which is a dichotomous variable indicating whether an individual has tested positive for diabetes.

Classification can also be used to distinguish healthcare fraud vs. non-fraud. Analytics applications that attempt to detect fraud and abuse often establish norms and then identify unusual or abnormal patterns of claims by physicians, laboratories, clinics, or others. Among other things, these applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. Analytics has taken the mass of data generated by millions of prescriptions, operations and treatment courses to identify unusual patterns and uncover fraud.

5). LOGISTIC ANALYSIS

In SAS/STAT, discriminant analysis and logistic are the primary means of classification. The logistic procedure will predict the values that can then be compared to accuracy. PROC LOGISTIC is performed to investigate the relationship between variables and responses in the dataset, to find out how certain variables are associated with the onset of diabetes, to interpret findings and prepare recommendations. The strength of a logistic regression is measured by the odds ratios, the receiver operating curve, and the p-value. A US healthcare provider created a predictive model for hospitalizations in their Cardiac and Asthma patients. They collated the information that they had about the patients to see what factors led to hospitalizations. When similar conditions occurred in other patients they took early intervention and reduced hospitalizations across these groups by over 80%.

6). ASSOCIATION ANALYSIS

In association analysis, the objective is to determine which variables go together. Such information can be useful for investigating associative relationships in healthcare.

Association Analysis is a technique used to identify events that occur together, possibly in a particular sequence or order. It can also be used to identify disassociations (i.e. events that do not happen after a particular one has already occurred).
Association Analysis can be used to answer questions such as
Which drug combinations are associated with adverse events?

Which of these associations are strong enough to need investigating?

Does the drug cause an adverse event or is the drug a common treatment for the condition?

PROC CORR can be conducted to see if the values of numeric variables are associated. Correlation coefficient contains information on both the strength and direction of the association. If measurements are strongly associated, it is expected to have a correlation value close to 1 or -1 if inversely associated. In contrast, if the measurements are less associated, a correlation value would be reduced to 0. Pearson correlation is a parametric measure of a linear relationship between two variables.

```
ods graphics on;
proc corr data=patient plots=matrix(histogram);
run;
ods graphics off;
```

The PLOTS=MATRIX(HISTOGRAM) option in the CORR procedure displays a symmetric matrix plot for all numeric variables in dataset .The histograms for these analysis variables are also displayed on the diagonal of the matrix plot. When the relationship between two variables is nonlinear or when outliers are present, the correlation coefficient might incorrectly estimate the strength of the relationship. Plotting the data enables to verify the linear relationship and to identify the potential outliers.

7). SAFETY SURVEILLANCE

All drugs have risks, and many are serious. Drugs are approved because their benefits are deemed to outweigh their risks. Rare but serious adverse events associated with vaccines or drugs are often nearly impossible to detect before the introduction of the agent in large populations. Participants observed that, with the increasing availability of electronic health data, opportunities have emerged to more accurately characterize and confirm potential safety issues. Studies in larger populations, with real-world dosing, longer duration of exposure, long-term follow-up, and comparison data based on current physicians' practices, are the most informative approaches to monitoring device and drug safety.
Ensuring that drugs have an acceptable safety profile and are used safely is a major public health priority. Safety surveillance is to assess and quantify known or suspected drug safety issues, identify and characterise potential new risks following product marketing, and to monitor product-use patterns. ICD-9-CM codes (flagged codes) are consistently more likely to indicate AEs, ICD-9-CM codes are best suited to targeted AE surveillance. PROC FREQ and INDEX() can search and identify AEs of interest. Statistical analysis is usually conducted by PROC FREQ with CHISQ options.

Poisson regression is a useful strategy when a small numbers of events because it does not depend on asymptotic results. The GENMOD procedure provides Poisson regression when specify DIST=POISSON and LINK=LOG in the MODEL statement.

```
proc genmod;
  class treatment agegroup gender race ;
  model counts = treatment agegroup gender race /dist =poisson offset = logrisk type3;
  estimate 'treatment' treatment 1 /exp;
  exact treatment / estimate = odds cltype = exact;
run;.
```

Data mining has a major impact on business and finance. Worldwide, all types of organizations are achieving measurable payoffs from this technology. Prescriber & patient segmentation and profiling can create actionable customer segments based on behavior observed through analysis of de-identified, patient-level healthcare data enhanced with consumer demographic and lifestyle behavior information.

8). QUANTILE REGRESSION ANALYSIS

Healthcare research data are not well distributed. They are often characterized by a high level of skewness and heterogeneous variances. A common healthcare expenditure model, for example, is that a small percentage of a population accounts for a large percentage of healthcare expenditures. Ordinary least squares (OLS) may not be able to correctly identify key drivers for the events under investigation. While quantile regression is a powerful tool to tease out heterogeneity and to compare patterns by quantile group. Using healthcare claims data, QR analysis can identified the end-of-life care, acute and post-acute care, and primary care are responsible for the expenditures for high-end users.

QR provides robustness and efficient estimates under non-Gaussian conditions.Healthcare influenced heterogeneously by many factors, are naturally suited for QR analysis. To conduct the quantile regression in SAS, on can perform QUANTREG procedure.

```
PROC QUANTREG DATA = sas-data-set <options>;
  BY variables;
  CLASS variables;
  MODEL response = independents / quantile = 0.05 to 0.95 by 0.1 seed=12345
plot=quantplot <options>;
RUN;
```

Health care expenditures are the area important to policy that is amenable to an analytical strategy that measures differences across the distribution. The average user of health care is obviously very different from the heavy user in terms of health status and clinical characteristics, but what about other factors such as race/ethnicity, gender, employment, insurance status, sociodemographic factors and other factors of policy interest? Quantile regression allows for analysis of these other differences that exist among heavy health care users in a way that is not possible with commonly used regression methods.

Many opportunities for using quantile regression exist in the health services literature. For example, in an article describing quantile regression methods to determine whether the weight (obese vs. overweight), dietary predictors of HbA1c levels and diabetics status (non-diabetics, Type I or II diabetics) are associated with the upper tails of glucose levels. QR technique can be a valuable tool to provide extra insights for both model construction and model diagnosis. Investigators should put this valuable tool to the research daily toolbox and pay more attention to the data distributions.

CONCLUSION

SAS analytic solutions can achieve business goals in healthcare related to cost control, revenue generation and strategic performance management. SAS analytic tools are used to explore clinical outcomes and risk tolerances to improve the overall quality of patient care.

Analytics applications in healthcare can have tremendous potential and usefulness. However, the success of healthcare analytics replies on availability of clean healthcare data. In this respect, it is critical that the healthcare industry consider how data can be better captured, stored, prepared, and mined. These analyses are further enhanced by adding information from other data streams, including physicians' notes, laboratories, and even consumer characteristics, to heighten the value and precision of the study. It is necessary to also explore the use of text mining to expand the scope and nature of what healthcare analytics can currently do. In particular, it is useful to be able to integrate data and text mining.

Analytics applications can be developed in the healthcare industry to determine preferences, usage patterns, individuals' current and future needs to improve their level of satisfaction. Predicting patient behavior is a natural progression of patient segmentation. By applying the knowledge from our segmentation and profiling analyses, we will be able to predict consumer behaviors, project trends, and forecast products, and markets based on unique segments such as ethnicities, cohorts, co-morbidities, and payer types.

Data Mining provides powerful tools that can reveal complex and hidden relationships in large amounts of data. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that help increase revenues, reduce costs and predict future results.

CONTACT INFORMATION

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

XXXXXXi, XXXXXX
XXXXX Pharmaceuticals
XXXX
XXX, PA XXXXX
Work Phone: XXX-XXX-XXXX
E-mail: XXXX@XXX.com

APPENDIX

```
%macro check_masking  (lib=, outdsn=);
proc contents data=&lib.._all_ out=memname (keep=memname) noprint;
run;

proc sort data=memname nodupkey;
  by memname;
run;

data _null_;
  set memname;
  call symputx( 'mem'||left(_n_), memname);
  call symputx( 'nmem', _n_);
run;

data maskname;
  set _null_;
run;

%do j=1 %to &nmem; ;
%put ***************** J=&j DATA =&&mem&j  ************************;
proc contents data=&lib..&&mem&j out=varname (keep=name  nobs) noprint;
run;

data _null_;
  set varname;
  call symputx( 'var'||left(_n_), name);
  call symputx( 'nvar', strip(_n_));
  call symputx( 'nobs', strip(nobs));
run;

data mask;
  set _null_;
run;

%do i=1 %to &nvar; ;
%put ***************** I=&i VAR=&&var&i  ************************;
  data temp1;
    set &lib..&&mem&j ;
        p=verify ( upcase(&&var&i), 'X .') ;
        if p>0 then p=1;
        keep p;
  run;

  proc means data=temp1 nway noprint;
    var p;
        output out=temp2 sum=sum;
  run;

  data temp3;
    set temp2;
        length result $200 ratio $50;
```

```
        if sum=0 then result = "&&var&i = Null values -- missing or de-identified by X";
    else result = "&&var&i = Valid values";
        ratio =strip(sum)||' / '||strip("&nobs");
        percent=sum*100/&nobs;
        format percent 6.1 ;
  run;

  data mask;
    set mask temp3 ;
  run;
 %end;

%let table=%scan(&&mem&j, 2,.);
data mask;
  set mask;
  length table $50;
  table =propcase("&&mem&j");
  Variable = strip( scan(result, 1, '='));
  Description= strip( scan(result, 2, '='));
run;

title "**================ J=&j DATA=&&mem&j ================***";
proc sort data=mask ;
   by sum result ;
run;

proc print data=mask;
  var Table variable description ratio percent;
run;

data maskname;
  set maskname mask;
run;
%end;

data &outdsn;
  set maskname;
run;
%mend;


%check_masking  (lib=xdata, outdsn=varchk);
```

Output Results:

Obs	TABLE	VARIABLE	DESCRIPTION	RATIO	PERCENT
1	Patient	FIRSTNAME	Null values -- missing or de-identified by X	0 / 14010	0.0
2	Patient	LASTNAME	Null values -- missing or de-identified by X	0 / 14010	0.0
3	Patient	MEDICALRECORDNUMBER	Null values -- missing or de-identified by X	0 / 14010	0.0
4	Patient	GENDER	Valid values	14010 / 14010	100.0
5	Patient	ISACTIVE	Valid values	14010 / 14010	100.0
6	Patient	PRIMARYKEY	Valid values	14010 / 14010	100.0
7	Patient	RACE	Valid values	14010 / 14010	100.0
8	Patient	AGE	Valid values	14010 / 14010	100.0