Paper 2240

Member Level Forecasting Using SAS Enterprise Guide® and SAS Forecast Studio®

Prudhvidhar R Perati, Leigh McCormack, BlueCross BlueShield of Tennessee, Inc., an Independent Company of BlueCross BlueShield Association

ABSTRACT

The need to measure slight changes in healthcare cost and utilization patterns over time is vital in predictive modeling, forecasting, and other advanced analytics. At BlueCross BlueShield of Tennessee, Inc., an independent company of BlueCross BlueShield Association, a method for developing member level forecasts creates a better way of identifying these changes across various time spans. The goal is to create multiple metrics at the member level that will indicate when an individual is seeking more or less medical or pharmacy services. Significant increases or decreases in utilization and cost are used to predict the likelihood of acquiring certain conditions, seeking services at particular facilities, and self-engaging in health and wellness. Data setup and compilation consists of calculating a member's eligibility with the health plan and then aggregating cost and utilization of particular services (i.e., primary care physician visits, pharmacy costs, emergency room visits). The computing power and complexity needed to execute monthly forecasts for over two million people for multiple measures requires in-database processing and various macro processes. SAS Enterprise Guide® is used to structure the data and SAS Forecast Studio® is used to forecast cost and utilization at a member level. Data is stored so that each metric is appended on a monthly basis. The insight provided by this member level forecasting methodology replaces subjective methods that used arbitrary thresholds of change to measure differences in cost and utilization.

Keywords: Base SAS®, SAS Enterprise Guide®, SAS Forecast Studio®, SAS/SQL®, SAS/MACROS®, In-Database Processing, Forecasting, Automatic Forecasting

INTRODUCTION

An efficient SAS® program that can generate monthly forecasts of cost and utilization for eligible members adds great value in analytics and ultimately data-driven business decisions. These SAS® methods and/or products are used in data acquisition, data cleaning, forecasting and creating a data mart to store and organize analytic outputs.

Explicit SAS/SQL® pass through methods were used to extract and summarize transactional data into a time series data format. Data were cleaned to adjust for appropriate business and forecasting criterion, such as missing data associated with a lack of health insurance coverage. A SAS® program constructed in SAS Forecast Studio® was incorporated to generate forecasts in a batch environment. With the advent of SAS Forecast Studio®, an analyst can choose between the best forecasting methods for the time series and produce more accurate, large-scale forecasts more frequently.

Twenty-four months of data were obtained for each member, and cost and utilization metrics summed by each month. At most, members were allowed to have two months of missing data. Based on the distribution of their time series data, each member receives a set of models to fit to their data and the model with the best mean absolute error (MAE) statistic is selected for forecasting. This paper gives a detailed overview of health care cost and utilization data as well as the methods and application of the member level forecast output.

HEALTHCARE DATA

A member's eligibility with a health plan is captured in a format with each unique enrollment being documented with an effective and termination date. This allows for knowledge of which dates a member has coverage with the plan as well as how long a member is "eligible" in a given time span. Administrative claims data consists of a member's interaction with the plan, such office visits or filling a pharmaceutical script. Data contained on these claims consists of place and date of service, provider information, amount

charged and paid for the services rendered, and specific diagnosis received and procedures received during that encounter.

There are multiple characteristics of healthcare data that are important to consider for time series analysis:

- Healthcare data is transactional in nature; a member's use of healthcare services are recorded in claims data at multiple times stamps which may are may not be evenly spaced. To fit the time series assumptions of evenly spaced intervals and for the ease of interpreting the forecasts, data are accumulated to a monthly time interval. An illustration on how transactional data (shown in Figure 1) is summarized to monthly time series (shown in Figure 2) is given below:

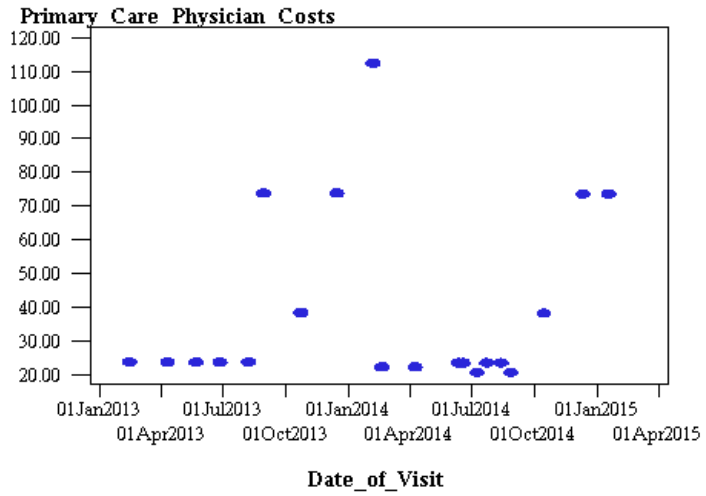

Figure 1: Transactional data of primary care physician costs for a member

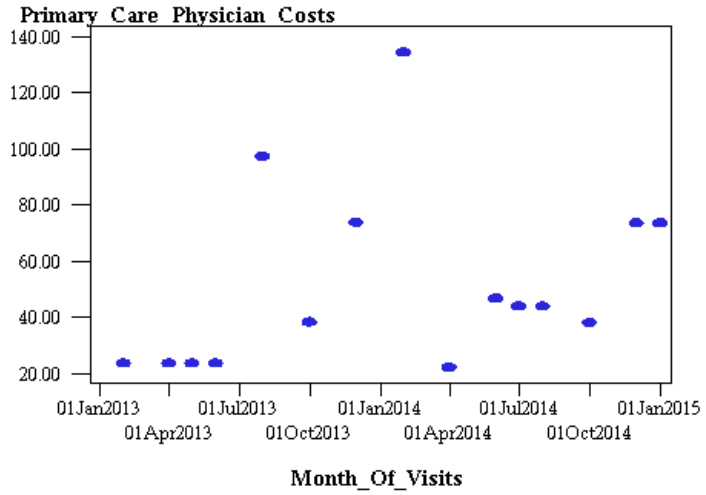

Figure 2: Monthly-time series data for primary care physician costs for a member

- Healthcare claims have a processing lag; once a member utilizes a healthcare service his/her claim is filed and series of events trigger to validate and process the claim. These events often

take time and the claims are not available in the data warehouse until perhaps weeks after the service occurred. When performing analysis using these data it is important to consider this lag in processing. Shown below in Figure 3 is a monthly summarized primary care physician costs obtained in early February 2015 until the end of January 2015 for a sample (n=100,000) of members.

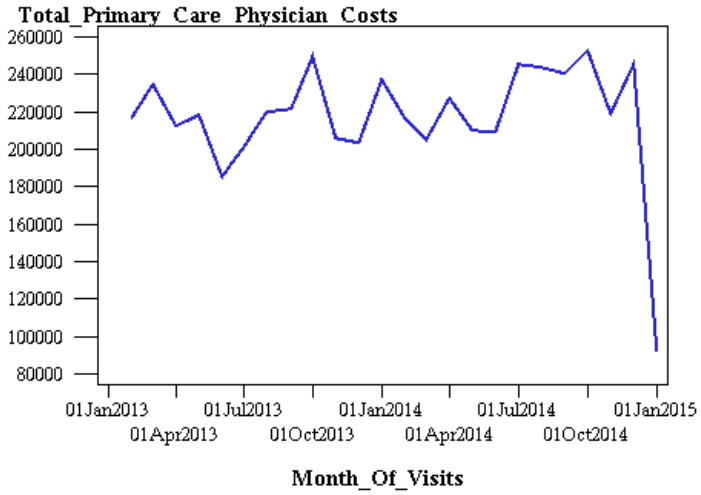

Figure 3: Monthly primary care physician costs for a random sample of members

A sudden decrease in costs is evident in the final month which is caused by the lag in processing of claims. Therefore data from the last month is not considered into this member level forecasting.

- Healthcare data are often intermittent (interrupted) in nature; when data is summarized at a monthly level, there are multiple months in a study period where the member would not have any encounters for certain types of service. Cost and Utilization for these months is denoted as zero. Figure 4 shows an example of this with zero values appearing for multiple months in the study period.

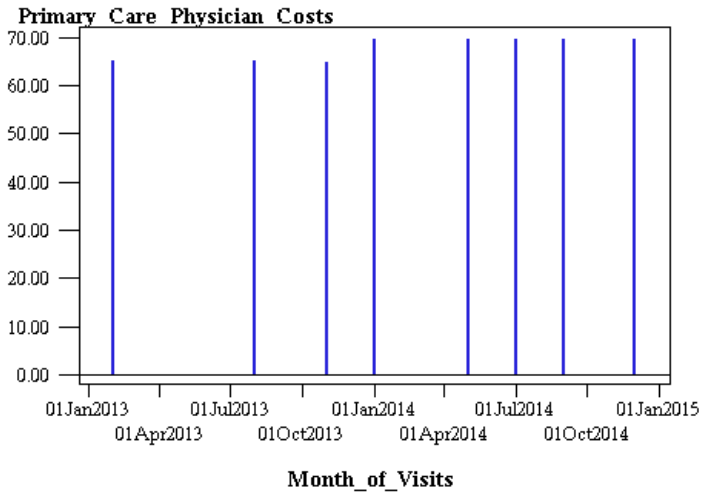

Figure 4: Monthly primary care physician costs for a member

Forecasting intermittent data poses various hurdles. SAS Forecast Studio® is equipped to forecast intermittent data. Background and methods of forecasting intermittent data using forecast studio were discussed by Michael Leonard in 2008 SAS proceedings[1].

All of these unique attributes of administrative claims data are dealt with using methods discussed below, that allows analysis to convey more accurate representations of member level changes in cost and utilization.

METHODOLOGY:

Data Acquisition:

A member can lose and obtain healthcare coverage many times throughout a given period of time. This leaves what could be considered "gaps" in coverage where data is unavailable for that member. For this exercise, members having coverage at the end of the study period and twenty-four months of continuous coverage with no more than a 60 day gap were considered.

Once the eligible population is determined, claims data for approximately two million members were extracted for a span of 24 months from a Teradata database. In-database processes were used to summarize cost and utilization data at the member-month level. Base SAS® features of SAS Enterprise Guide® were used to write SAS/SQL® explicit pass through queries to extract data from the Teradata environment. Additional logic was written to categorize utilization (i.e., emergency room versus primary care office visits), summarize the data at the member-month level and deliver only the summarized data back to SAS®, whereby saving valuable I/O time. Claims data cleaning steps such as adjusting for voided claims (i.e., those claims that are retroactively edited) and providing a time lag consideration for the claims to be processed are accounted for with in this SAS/SQL® explicit pass through query to Teradata. SAS/Macro® language was used to relay the dates to import new data on a monthly basis.

The following cost and utilization variables are summarized to a member-month level

Utilization	Costs
Primary Care Physician Visits	Primary Care Physician Costs
Inpatient Visits	Inpatient Costs
Emergency Room Visits	Emergency Room Costs
Outpatient Visits	Outpatient Costs
Specialist Visits	Specialist Costs
Ancillary Visits (i.e., laboratory services)	Ancillary Costs
Pharmacy Scripts	Pharmacy Costs

Table 1: Cost and utilization variables for forecasting

Data Attributes:

As with all healthcare data, there can be an abundance of missing values. These missing values can represent different things depending on how they are derived. For instance, structural missing values, or structural zeroes, occur when a member is eligible during a month and did not have a healthcare encounter or incur cost for a particular type of service.

Missing values occur during a period when a member has a gap in coverage and data points are unknown for that particular stint in a member's history. Missing values are found for less than one percent (~0.56%) of the observations and are often imputed with mean, median, mode or other appropriate statistics.

Figure 5 illustrates the distributions of the number of primary care physician visits in a month. It is evident that the mode for the primary care physician visits is zero and thus is the case for the other cost and utilization variables listed in Table 1.

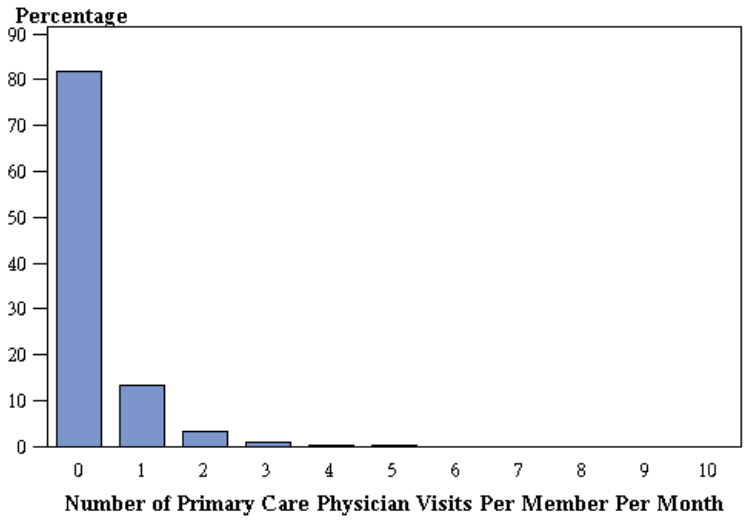

Figure 5: Frequency distribution of primary care physician visits

Taking the missing value percentage into consideration and the mode of the data, it was ideal to replace missing data with zeros (mode of the data).

Cases do exist where a member has not utilized any of the healthcare services in the study period and the values of cost and utilization for these members would be all zeros (zeros for each month throughout the study period). Since all the values are zeros, forecasts would not be generated for these members, and the future forecasts would be zero. Hence removing these members prior to forecasting would save valuable computing time and resources.

Once the accumulated data are imported from the database and the missing values are imputed and cases where there are no events are dropped, data are now ready for forecasting.

Forecasting:

After importing the data to SAS Forecast Studio® and the necessary options are chosen to forecast the dependent variable by member for a span of two months, diagnostics are automatically performed on each time series for trend, seasonality, linearity/nonlinearity (appropriate transformation methods), and intermittency. Once diagnostics are performed, SAS Forecast Studio® associates the appropriate group of models to forecast the time series data. Models are fit and evaluated on the full range of data. In a case where multiple models are fit, the model with the lowest MAE is selected as the champion model. List of possible model methods include:

- ARIMA
- Subset (factored) ARIMA
- Unobserved Components
- Exponential Smoothing
- Intermittent Demand
- Multiple Regression
- Moving Average

- Curve Fitting
- Random Walk

Models were occasionally combined when the combining models generated better forecasts than the individual models alone.

A detailed explanation of the use of SAS Forecast Studio® is beyond the scope of this paper. SAS Forecast Studio® documentation gives a detailed explanation of the uses of the software and for a quick reference review the paper from Brenda Wolfe on "Introducing SAS Forecast Studio®".[2]

Once the forecasts are complete, the forecasting project is closed and the SAS® code titled CREATE_PROJECT_IMPORT_DATA is obtained from the location where the forecasting project is saved. Using this code, the required forecasted outputs at the member level can be generated.

The entire process flows as follows: in-database processing is used to extract, clean and summarize the data, CREATE_PROJECT_IMPORT_DATA code is used for forecasting each cost and utilization variable and since forecasting each cost and utilization variable is an independent event, RSUBMIT functionality of SAS/CONNECT® is used to process the forecasting code for each cost and utilization variable in parallel fashion.

SAS® code combining the above mentioned steps, SAS/MACRO® processes, and a base SAS® program to append the forecasted datasets to the data mart are combined to generate the consolidated SAS® program. This program is executed monthly in a batch mode (uninterrupted fashion) using the UNIX "at" functionality.

RESULTS AND PERFORMANCE

Results and discussion are provided for data pertaining to primary care physician visits. Figure 6 illustrates the distribution of selected models for this metric. The majority (~70%) of members' data was fit utilizing the Intermittent Demand Model (IDM).

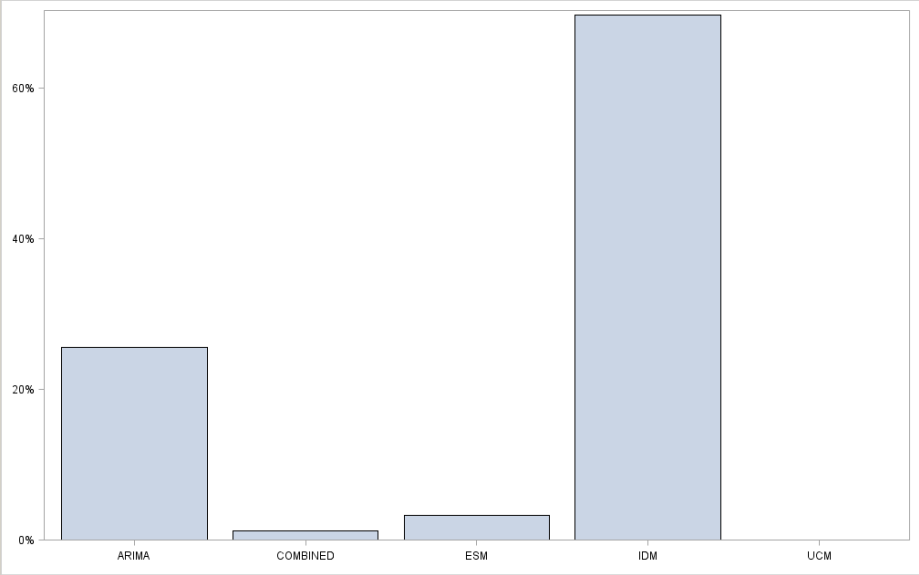

Figure 6: Distribution of models selected for primary care physician visits forecasts

Large scale forecasting at the member level results in multiple metrics and statistics of fit for each model. Along with the fit statistic such as MAE/MAPE, diagnosing each forecast individually by looking at the demand curves for Intermittent Demand Models and auto-correlation plots for ARIMA models gives a great insight into the performance of a forecast. However, this is a tedious process and highly inefficient to manage for millions of forecasts on a monthly basis.

With the availability of data from the forecasted period, an analyst can compare the actual, observed data to the forecasts and compare the efficiency of forecasts. The observed data was compared with forecasts from the High Performance Forecasting (HPF) of SAS Forecast Studio®, with naïve forecasts of:

- Change Models: Current forecast is same as the previous month's forecast
- Seasonal Models: Current forecast is same as the observed value for this month from the previous year
- Average Models: Current forecast is same as the average of observed measures from the historical 24 months of data

The frequencies of the methods with the lowest error are shown in Figure 7.

Forecasting Method	Percent
HPF Forecasting	71.45
Change Models	12.99
Seasonal Models	11.10
Average Models	4.46

Figure 7: Comparison of HPF forecasting to naïve forecasting methods

While using various methods in predicting primary care physician costs, it has been observed that for seventy-one percent of the series SAS Forecast Studio's® High Performance Forecasting methods have the lowest error compared to the other naïve methods in predicting the events of costs and utilization.

Predicted and actual values of ten members are shown in Table 2 below. In general, the predicted visit counts are closely matching to the actual observed visit counts. There would be cases when other models, such as the average models have a lower error value compared to the HPF predictions. However, relying on forecasting predictions would typically yield a higher percentage of lower error values.

Member Number	Predicted PCP Visits	Actual PCP Visits	Average PCP Visits	Predicted PCP Costs	Actual PCP Costs	Average PCP Costs
M1	0.12	0	0.22	$4.20	$0	$15.24
M2	2.36	2	2.26	$660.95	$175.7	$770.11
M3	0.59	1	0.39	$22.87	$270.88	$51.98
M4	0.13	0	0.17	$19.62	$0	$27.93
M5	0.84	0	0.22	$69.34	$0	$45.85
M6	0.27	0	0.13	$97.97	$0	$58.47
M7	0.19	0	0.17	$28.97	$0	$55.32
M8	0.17	0	0.22	$20.86	$0	$23.41
M9	0.17	0	0.22	$12.37	$0	$15.91
M10	0.28	0	0.30	$18.74	$0	$42.41

Table 2: Actual, predicted and average values for primary care physician visits and costs.

With champion models selected and output generated on a monthly basis for each member and each cost and utilization metric, input into subsequent analytics begins to test significance in predicting members' healthcare related behaviors. These 14 sophisticated metrics help monitor and convey trend for certain service types, and detect slighter differences amongst member that more subjective methods

may have missed. All-in-all, this approach provides greater insights into a diverse member population and allows more accurate and detailed data-driven recommendations.

CONCLUSION

Member level predictions of cost and utilization can be of value in various predictive models, advanced analytics and in stratifications where it is essential in accounting for member differences in trend. Features offered by SAS Forecast Studio® can be efficiently combined with SAS® macro processes, in-database processing techniques and SAS Connect® techniques to create efficient SAS® programs that can be run in a batch mode to update a data mart with forecasts of cost and utilization. Current SAS® process for forecasting takes around 18-19 hours. Since the forecasting is performed at a member level for around two million members, certain major time taking processes cannot be avoided. SAS® in-database procedures for high performance automatic forecasting utilizing the parallelism of Teradata can improve the performance and reduce computing time tremendously.

REFERENCES

[1] Leonard, Michael., Elsheimer, Bruce., John, Meredith., and Sglavo Udo. (2008),"Small Improvements Causing Substantial Savings - Forecasting Intermittent Demand Data Using SAS® Forecast Server" Cary, North Carolina: SAS Institute, Inc.

Available at https://support.sas.com/resources/papers/sgf2008/forecasting.pdf

[2] Wolfe, Brenda., Leonard, Michael., and Fahey, Paddy. (2005), "Introducing SAS® Forecast Studio" Cary, North Carolina: SAS Institute, Inc.

Available at http://www2.sas.com/proceedings/sugi30/193-30.pdf

CONTACT INFORMATION

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

Prudhvidhar Reddy Perati
prudhvidhar_perati@bcbst.com

ACKNOWLEDGEMENTS

Special thanks to Brandon Cosley, PhD for valuable suggestions and edits regarding this methodology and resulting paper.