Paper PH-26

Identifying Superutilizers in the Medicaid Population—A Simple and Powerful technique

Aran Canes, Cigna Corporation

ABSTRACT

The purpose of this study is to test whether we can predict high cost recipients in the Medicaid population one year in the future with a high degree of certainty using data from the recipients' claim histories and relatively simple predictive modeling methods.

We use logistic regression as the modeling tool and then refine the initial approach by binning continuous predictors and testing whether by clustering or by calculating Spearman and Hoeffding correlation coefficients we can eliminate irrelevant variables. We are able to show that a significant number (over 2,000) of these recipients can be predicted. Also, the positive predictive value for these recipients is close to 90%. This high positive predictive value means that resources do not need to be expended on recipients who are not likely to be true super-utilizers.

This is an important finding, as states which do not have the internal resources to pay for vendor predictive modelling software, or develop their own complex predictive model, can significantly improve on using only provider referrals to identify likely high-cost recipients. It is hoped that this study can be a reference to such states in using data analytics to guide their search for recipients who are in need of care management.

INTRODUCTION

Predicting high-cost recipients is a priority for Medicaid programs around the country. It is estimated that the top 5% of patients account for more than 50% of the total cost.[1]

It is also well-known that patients with multiple co-morbidities could benefit from coordinated care.[2] Therefore, it is important both for cost and quality of care concerns to identify likely future high cost recipients.

Because of this, commercial software companies sell predictive modeling software to Medicaid programs around the country. Some states, on the other hand, have chosen to build their own predictive models.[3] There are costs and benefits to each approach. Commercial software is expensive and may not be tailored to the particularities of a given state's Medicaid population. On the other hand, statisticians, computer programmers and medical experts must be hired to build an internal predictive model, and this can be resource and time intensive. There is not agreement, to my knowledge, among Medicaid Directors on which approach is better. One prominent doctor is even quoted as saying,

"For all the stupid, expensive, predictive-modelling software that the big venders sell, you just ask the doctors, 'Who are your most difficult patients?,' and they can identify them. [4]

In this paper we offer another alternative, a model that uses Proc Logistic in SAS® to generate predictive probabilities regarding who is likely to be high cost. The predictor variables within the model are all derived from claims data and consist simply of prior cost, diagnoses and utilization statistics. By calculating evaluation statistics such as the positive predictive value, the model proves not only simple and easy to implement but also a highly effective means of prediction. We believe it offers a simple and powerful technique to identify likely future superutilizers that could be employed by state Medicaid agencies who do not want to make the time or money investments of the traditional approaches.

METHODOLOGY

Initially, our data population included all 2,069,035 recipients who were eligible for Medicaid as of March 31, 2013. Because of the different nature of health problems among children, adults and senior citizens we divided this population into three distinct age categories (0-20, 21-64 and 65 and above). We excluded recipients who were also eligible for Medicare as Medicaid only pays part of their medical costs. Because the vast majority of senior citizens have Medicare, we built two different models: one for recipients between the ages of 0-20 and one for recipients 21-64. For space reasons, this paper will discuss the findings for recipients between the ages of 0-20 only.

The next decisions were the choices of predictive model and a threshold for determining what constitutes a high-cost recipient. For purposes of this study we were not so much looking to predict how much a recipient would spend, but rather whether the recipient is likely to be costly enough to warrant managing their care. After discussion with medical experts we determined that if a recipient would cost the Medicaid program more than $5,000 per quarter, we would want medical experts to examine that person's claims history to determine if their care is adequate. Since we now had a binary outcome and wanted a probability of each recipient being high-cost in the future, a logistic regression seemed the tool of choice.

We then computed statistics on a recipient/quarterly basis: that is, each recipient had their own value of each predictor variable for every quarter of 2011 and 2012. The predictor variables were of three different categories: prior costs, diagnoses and utilization.

We separated cost variables into total physical, behavioral and long-term care costs. We then calculated utilization statistics by calculating the number of inpatient admissions, inpatient readmissions and ER visits per quarter per recipient. Diagnosis predictor variables were incorporated into the model by summarizing all ICD-9 diagnoses into one of 110 different diagnostic categories. These diagnostic categories were developed at the University of California, San Diego and are known as Chronic Illness and Disability Payment System (CDPS) categories.[5]

We had data from each quarter of 2011 and 2012, including total cost per recipient, the cost associated with physical, behavioral and long-term care costs, how many inpatient admissions, readmissions and ER visits each recipient had, and whether the recipient had one of 110 diagnostic categories for at least one claim/encounter in each quarter.

The next step was to determine coefficients for each of the predictor variables by using some of the data as a training data set. Because the Commonwealth of Pennsylvania experiences an approximate nine month lag between when it receives claims and the date of service, we used data from 2011-Q1 and 2011-Q2 to predict recipients who would be high cost in 2012-Q2. The following is the SAS code from Proc Logistic which generated the coefficients on the predictor variables.

```
PROC LOGISITIC DATA=kidsnondualcompare;
       MODEL highcost3a (EVENT='1')=
             phpriorbinned bhpriorbinned ltcpriorbinned…
            (other predictor variables)
       STB CLODDS=pl RIDGING=none;
       SCORE DATA=kidsnondualcompare3 OUT=kids300013Q4;
run;
```

We next sought a validation data set to test how well these predictors correctly identified who would be high cost. To do this, we used data from 2011-Q2 and 2011-Q3 to predict who would be high cost in 2012-Q3. Finally, we used data from 2012-Q3 and 2012-Q4 to predict which recipients would be high cost in 2013-Q4 once the model had been validated.

Finally, we had to establish a probability threshold for determining whether the recipient was likely enough to be high cost in one year's time that they should be selected for potential management of their health care needs. Because it is resource intensive to manage these recipients, we determined that a high threshold was appropriate. A 90% probability that a recipient will be high cost is high enough that one can be almost certain that these recipients should be included, but is low enough to be sure that a substantial number of recipients will be included.

RESULTS

For a first effort, we tried the "kitchen sink" approach; that is, we used all utilization statistics, CDPS categories, and expenditure data, a total of 233 independent variables. Then we used 2011-Q1 and 2011-Q2 to establish the parameters of the model and tested it on 2012-Q3 recipients.

Unsurprisingly, the model did not meet expectations. There were a total of 2,871 predictions above the 90% threshold but the positive predicted value of these predictions was only 78%. That is, only 78% of these predictions were true positives when evaluated on the validation data set.

Predictions_Above_90%	Predictions_Above_90%_Matches	Positive Predicted Value (Matches/Total Predictions)
2,871	2,243	0.78

Table 1 Positive Predicted Value of First Effort

We then tried several different techniques to improve the model. The first was to test the linear relationship between the dependent variable and the continuous expenditure variables. In all types of linear regression the dependent variable should have a linear relationship with the predictor variables. Because health care costs tend to follow a gamma distribution, it is not surprising that, without binning cost, there was not a linear relationship between total cost and the dependent variable. One method of testing the linear relationship is to plot the empirical logit which is an estimate of the proportion of the target event within a binned continuous predictor variable.

The following is the SAS® code which generates a binned variable.

```
PROC RANK DATA=kidsnondualbinneda6 GROUPS=100 OUT=out;
       VAR ph_c_11q2;
       RANKS bin;
WHERE ph_c_11q2>5000;
run;


PROC MEANS DATA=out NOPRINT NWAY;
       CLASS bin;
       VAR ph_c_11q2;
       OUTPUT OUT=endpts MAX=MAX;
run;

FILENAME rank1 "c:\Predictive Part 2\rank1.sas";
DATA _null_;
FILE rank1;
SET endpts end=last;
       IF _n_=1 THEN PUT "select;";
       IF not last THEN DO;
              PUT " when (ph_c_11q2 <= " max ") phpriorbinned=" bin ";";
       END;
       ELSE IF last THEN DO;
       PIT " otherwise phpriorbinned= " bin ";";
       PUT "end;";
       END;
RUN;

DATA kidsnondualbinneda6test;
SET kidsnondualbinneda6;
       %INCLUDE rank1 / source;
RUN;
```

Chart 1 below shows the empirical logit versus unbinned physical health expenditures from the second quarter of 2011. Although there is a roughly linear relationship, the pattern is not smooth.

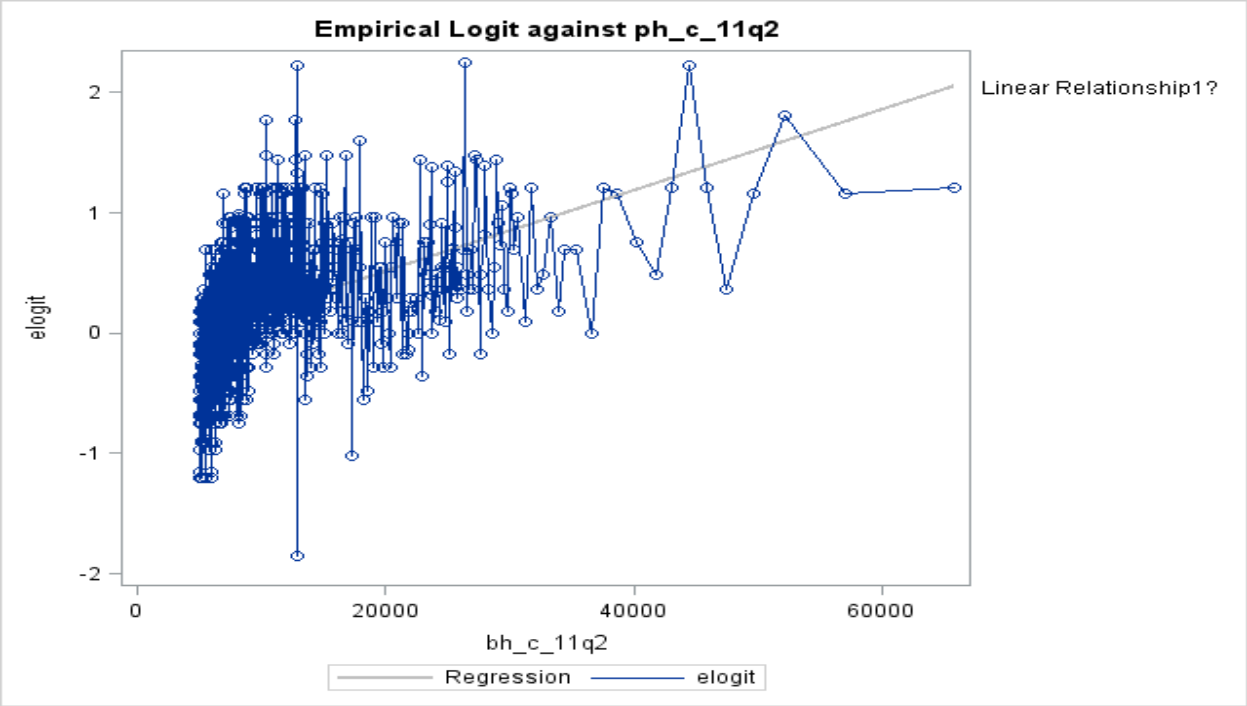

Chart 1 Empirical Logit vs. Physical Health from the Second Quarter of 2011

However, once we bin physical health into one hundred different categories the linear relationship is much more pronounced, as chart 2 below shows.

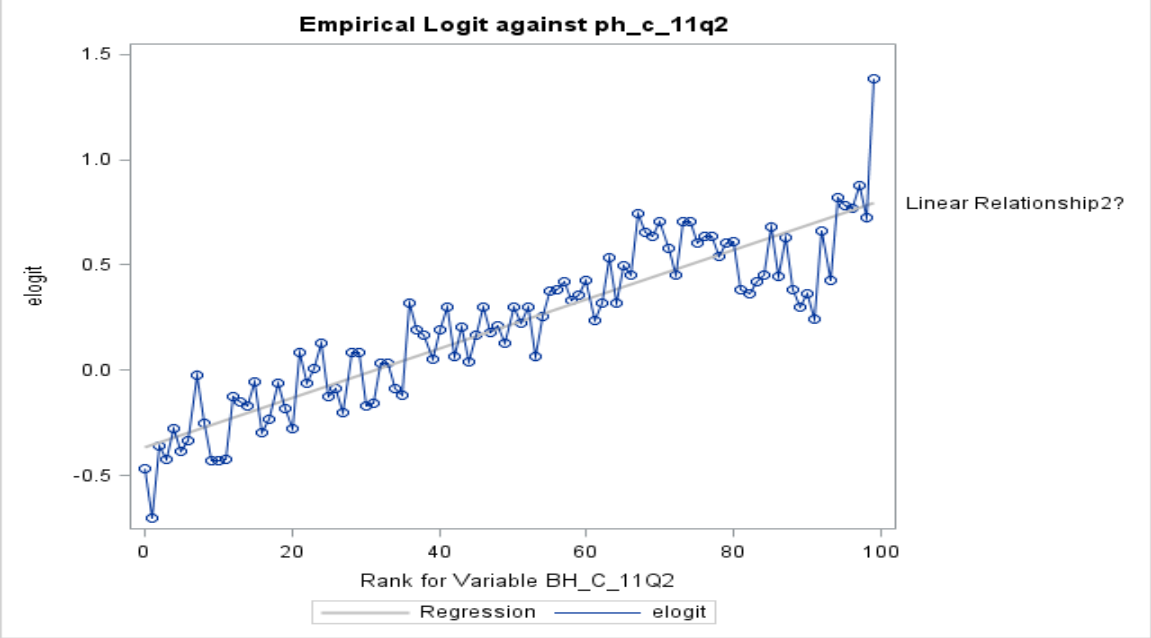

Chart 2 Empirical Logit vs. Binned Physical Health from the Second Quarter of 2011

The ultimate test of the success of binning expenditure variables is, however, whether they increased the predictive power of the model. Table 3 below shows the positive predicted value once all expenditure variables are binned in 100 categories instead of inputting all of them directly to the model. Despite a slight reduction in the overall number of correct predictions, the positive predictive value has increased greatly.

Predictions_Above_90%	Predictions_Above_90%_Matches	Positive Predicted Value
2,410	2,145	0.89

Table 2 Positive Predictive Value with Binned Expenditure Variables

Another technique which could result in increased predictive power is to cluster the independent variables.

Proc Varclus separates variables into clusters by starting with all variables in a single cluster and then performing a principal components analysis. In simple terms, if there are multiple principal sources of variability the cluster is split until there are enough clusters that each has only one principal source of variability. The variable that has the most correlation with other variables in its cluster and the least correlation with other clusters is then selected as the representative variable.

The following is the SAS® code which generates clusters.

```
PROC VARCLUS DATA=cluster MAXEIGEN=1 SHORT HI;
       VAR priorbaby1 priorbaby2 …;
run;
```

In this case, Proc Varclus identifies 92 clusters for the 220 independent diagnostic variables. The representative variable is then selected from each cluster and the logistic regression is repeated with these 92 variables (plus the previously binned expenditure variables and utilization statistics).

Predictions_Above_90%	Predictions_Above_90%_Matches	Positive Predictive Value
2,707	2,358	0.87

Table 3 Positive Predictive Value with Clustering Algorithm

At first glance this model might look more successful than the previous model. More than 200 correct matches have been added. However the positive predictive value has fallen by 2% because approximately 300 additional recipients are predicted. The PPV on those added recipients is only 71%, reducing the overall PPV and below the threshold we seek. As a result, we do not include the clustering algorithm in the model.

A final technique to potentially improve the predictive power of the model is to eliminate variables which don't have a relationship with the target variable. One means of accomplishing this is to compute Spearman and Hoeffding Correlation coefficients. Spearman correlation coefficients identify whether there is a monotonic linear relationship between two variables. Hoeffding correlation coefficients measure a wider variety of linear associations between two variables.

The following code generates Spearman and Hoeffding Correlation coefficients between all the diagnostic predictor variables and the dependent variable.

```
PROC CORR DATA=cluster SPEARMAN HOEFFDING RANK;
       VAR priorbaby1 priorbaby2 …;
       WITH highcost3;
run;
```

If a variable has a p-value greater than 0.5 on both the Hoeffding and Spearman correlation test (giving 1 minus the probability that the null hypothesis of no correlation is falsified), that variable is likely an irrelevant variable and can be eliminated from the regression. In other words, if no linear or non-linear relationship is found between an independent variable and the dependent variable then the independent variable can be excluded from the final model.

Chart 3 below shows a scatter plot of the ranks, by highest degree of correlation with the dependent variable, of the Spearman and Hoeffding correlation statistics. All variables to the right of the vertical line and above the horizontal line do not have significant Spearman or Hoeffding correlation coefficients. As one can see, almost all the variables have some type of linear relationship with the dependent variable. Elimination of irrelevant variables is not a technique which will lead to a significantly higher number of predictions or a higher positive predictive value. Thus, calculation of the Spearman and Hoeffding correlation coefficients confirmed that all the diagnostic variables should be retained in the final model.

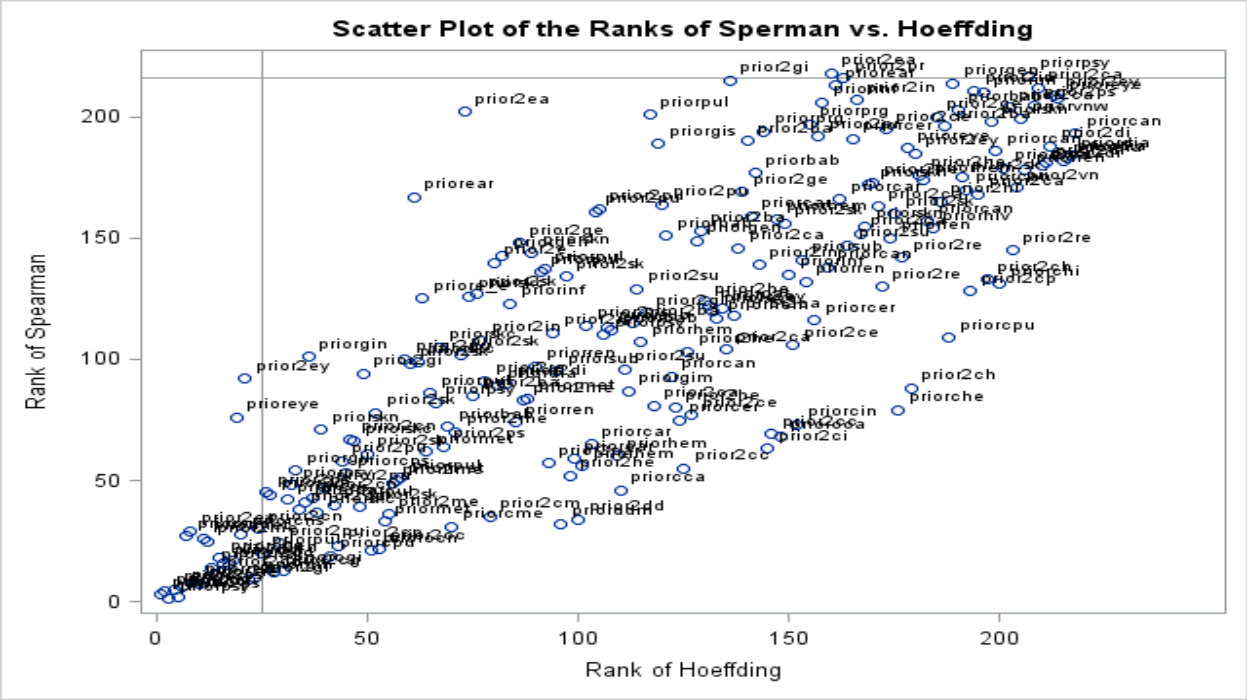

Chart 3 Scatter plot of Spearman vs. Hoeffding Correlation Coefficients

In sum, the most successful predictive model used binned expenditure variables, all diagnostic categories and some utilization statistics. The positive predictive value of this model was 89%, which is very close to the 90% threshold chosen prior to creating and testing the model.

BENCHMARKING

	Selected as High Cost	Truly High Cost	Percentage
Predictive Model	2,410	2,145	89%
Random Selection of Prior High Cost Recipients	2,410	1,060	44%
Highest Cost Members	2,410	1,590	67%

Table 4 Predictive Model vs. Benchmarks

What are some appropriate benchmarks to compare the predictive power of the model to? In the most successful model there were a total of 2,410 predictions, of which 2,145 were correct. What if we had simply selected 2,410 recipients who were high cost one year ago and selected them for potential care management? This would certainly be a simpler approach than building and applying a full logistic regression model. However, only 44% of the high cost recipients from one year ago repeated as high cost recipients one year later. Considering that the positive predictive value in the predictive model is 89% we have over double the match rate. What if we had chosen, however, the most expensive 2,410 recipients from one year ago? Only 1,590 of these repeated as high cost recipients one year later, resulting in a match rate of 66%. Thus, the predictive model was significantly more successful than simply predicting that the highest cost recipients would repeat.

CONCLUSION

The simple approach of predicting superutilizers in the Medicaid population by using a logistic regression on prior cost statistics, diagnostic categories and utilization statistics has been shown to be highly successful once the model has been adjusted by binning all prior expenditures. Using training and test datasets, one can predict high cost recipients at an almost 90% rate, which is high enough to warrant potentially managing their health care needs both to improve their quality of care and reduce costs.

ACKNOWLEDGMENTS

The author would like to thank Dr. David Kelley and Mrs. Holly Alexander for the opportunity to create this model.

CONTACT INFORMATION

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

Name: Aran Canes
Enterprise: Cigna
Address: 900 Cottage Grove Rd.
City, State ZIP: Bloomfield, CT 06002
Work Phone: 860-902-9876
E-mail: aran.canes@cigna.com

[1] Ash, Arlene, Zhao Yang, Ellis, Randall P., Kramer Marilyn Schein. December 2001. Finding Future High-cost Cases: Comparing Prior Cost Versus Diagnosis-based Methods. *Health Services Research*. 194-206. Chicago, IL: Health Research & Educational Trust.

[2] An example is the work of Jeffrey Brenner in Camden, NJ see:
Gawande, Atul. January 2011. The Hot Spotters. 41-53. *The New Yorker.* New York, NY: Conde Nast.

[3] The state of Washington provides an example. Their work is summarized in:
Knutson, Dave, Bella, Melanie, Llanos, Karen. Predictive Modeling: A Guide for State Medicaid Purchasers. Kaiser Permanente Community Benefit and Aetna Foundation. August 2009. Available at *http://www.chcs.org/publications3960/publications_show.htm?doc_id=992610*

[4] *The Hot Spotters*, ibid.

[5] Gilmer, Todd. University of California, San Diego Chronic Illness and Disability Payment System. University of California, San Diego. 2012. Available at http://cdps.ucsd.edu/