Paper SDA-04

Detecting Medicaid Data Anomalies Using Data Mining Techniques

Shenjun Zhu, Qiling Shi, Aran Canes,
AdvanceMed Corporation, Nashville, TN

ABSTRACT

The purpose of this study is to use statistical and data mining techniques in Base SAS(R) and SAS(R) Enterprise Miner™ to proactively reduce the number of false positives caused by data anomalies in Medicaid pharmacy claim data when employing a rule-based approach to identify overpayments. Typically rule-based techniques are based on specific state Medicaid laws and policies using certain formulas to detect and identify over charged payments. False positives are defined as an identified overpayment that is erroneously positive when a claim was paid correctly due to data anomalies or unknown factors. False positives substantially increase the amount of time and resources spent by the auditors. The specific objective of the study is to detect and reduce data anomalies by examining the relationships among key variables such as Medicaid amount paid (MAP), average wholesale price (AWP) and quantity of service in Medicaid pharmacy claim data.

Pharmacy claim data were simulated and the overpayment was calculated by a rule-based approach developed by AdvanceMed Corporation. Different data mining techniques such as the studentized residual, leverage, Cook's distance, DFFITS and clustering were utilized to capture the abnormal claims and reduce the number of false positives. The results of this analysis indicated that the clustering statistical method is the best approach to detect these kinds of data anomalies, followed by the DFFITS method.

INTRODUCTION

AdvanceMed specializes in helping healthcare organizations evaluate and assess the integrity of their health and pharmacy benefit programs. AdvanceMed conducts sophisticated data analysis to detect potential fraud cases from both the pre and post payment perspective using rule violations, statistical outliers, etc. to identify health care fraud and abuse. AdvanceMed aligns itself with cutting edge resources, developments, and capabilities which allows for progressive healthcare integrity in today's fluid environment. Through these efforts, AdvanceMed brings forth all the necessary elements to provide the client with the means to successfully meet its missions.[1]

METHODOLOGY

A simulation was conducted based on Medicaid data. Abnormal claims were added into the simulation data to test different data mining techniques used to detect data anomalies. Below is a rule-based calculation methodology used by AdvanceMed to detect the overpayment from pharmacy claim data. This rule-based algorithm is to identify overpayments where state Medicaid paid more drug units than state policy allowed.

If quantity of service (QOS) is greater than the maximum units (max units) permitted by the state, AdvanceMed can calculate the overpayment by the following formula:

$$\text{Overpayment} = \text{MAP} - (\text{AWP} * \text{discount\_rate} * \text{max\_units} + \text{dispense\_fee}). \quad (1)$$

The discount rate and dispensing fee are constants for a specific state. Hence by (1), we will have many false positives for identified overpayment if there exist abnormal claims related to MAP or AWP.

With the exception of strikeouts and errors, MAP should be calculated by a formula using AWP and QOS for each prescription. Below is a formula AdvanceMed uses to define the relationship between MAP and QOS if no other third party payment exists.

$$\text{MAP} = \text{AWP} * \text{QOS} * \text{discount\_rate} + \text{dispense\_fee}. \quad (2)$$

The discount rate and dispensing fee are constant for any prescribed prescription. We can infer from this equation that there is a linear relationship between MAP and the product of AWP and QOS. Hence we define a new variable called 'AQ' and let AQ=AWP*QOS. Then we perform the bivariate association analysis computing the Pearson correlation coefficients between MAP and AQ. In the simulated data the Pearson coefficient equals 0.91 which means there is a strong positive linear relationship between MAP and AQ. We then perform regression analysis predicting MAP from AQ.

Consider the linear regression model MAP= $\alpha_1 + \alpha_2 *$AQ + ε where the errors ε are independent and all have the same variance. Observations which have an extreme studentized residual or leverage for the fitted regression model can be identified as outliers. Cook's distance is a measurement of the influence of the i-th data point on all the other data points. The higher Cook's distance is the more influential the point is. We consider the claims when Cook's distance is greater than 4/n as outliers. DFFITS shows how influential a point is in a statistical regression. More specifically, it is the difference between the fitted (predicted) values calculated with and without the i-th observation. We identify the claims with DFFITS greater than 2*sqrt(k/n) as outliers (where k is the number of predictors and n is the number of observations).

Clustering is a statistical method of unsupervised learning. It puts a set of observations into subsets (called clusters) so that observations are clustered which have similar patterns between the variables. Since there are three distinct drugs in the table, we determined the number of clusters as k not less than three. SAS Enterprise Miner uses the clustering cubic criterion (CCC) cutoff value as its main criteria in the selection of number of clusters. In the average linkage method, the distance between two clusters is defined as the average of the distances between all pairs of objects, where each pair is made up of one object from each group. The segment identifier is assigned a role of segment. The cluster selects initial seeds that are very well-separated using a full replacement algorithm. The clustering methods in the Cluster node perform disjoint cluster analysis on the basis of Euclidean distances. SAS Enterprise Miner uses the Convergence Criterion Value property to specify the value of the convergence criterion in the computation of cluster seeds. The default convergence value is 0.0001.

RESULTS

The simulated pharmacy claim dataset consists of information about Medicaid pharmacy services. The response variable is the overpayment, calculated based on state policy. Possible explanatory variables include various measures of Medicaid pharmacy service. We add some aberrant records to the AWP in the simulated dataset to evaluate the effects of AWP data anomalies on the identified overpayments in the results. The data structure employed to calculate overpayment by a rule-based methodology is as below:

Table 1: Data Structure for Simulated Pharmacy Claim Table with Calculated Overpayment

Type of	Normal Claims	Abnormal Claims	Total
Total Observations	2,998	300	3,298
Number of Observations for Overpayments	63	9(False Positives)	72
Identified Claim Count Rate (%)	2.10%	3.00%	2.18%

The five highest and lowest overpayments for each drug are below:

Figure 1: The Five Highest and Lowest Overpayments for Each Drug

The UNIVARIATE Procedure
Variable: overpay

Extreme Observations

Lowest			Highest		
Value	national_drug_code	Obs	Value	national_drug_code	Obs
213.076	00156820	3227	2090.95	0067988	3294
214.272	00156820	3228	2092.27	0067988	3295
215.089	00156820	3229	2095.42	0067988	3296
220.137	00156820	3230	2097.28	0067988	3297
220.900	00156820	3231	2118.63	0067988	3298

We examined the regression command predicting MAP from AQ. We outputted several statistics that will be needed for the next few analyses as a dataset called “rx_res”. These statistics include the studentized residual (called r), leverage (called lev), Cook's Distance (called cd) and DFFITS (called dffit).

First, we used studentized residuals to identify outliers. The studentized residuals were retrieved from the previous regression analysis output. Ninety-two claims with studentized residuals either less than -1.42 or greater than 3 were identified as outliers (data anomalies).

Figure 2: Studentized Residuals Distribution

Variable: r (Studentized Residual without Current Obs)

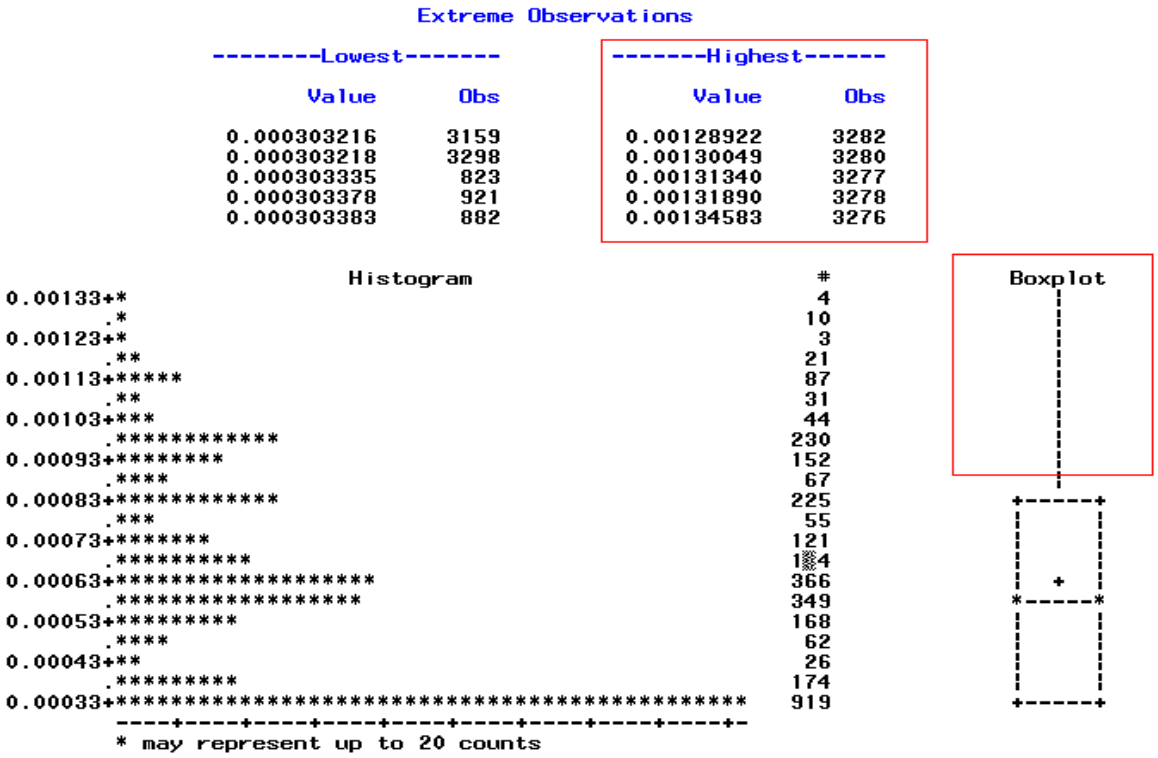
Extreme Observations

Lowest		Highest	
Value	Obs	Value	Obs
-1.72026	3270	3.63103	3177
-1.45163	3227	3.65464	3209
-1.44168	3228	3.66848	3133
-1.42836	3229	3.69934	3197
-1.42506	3230	3.79119	3165

```
Histogram                                  #    Boxplot
 3.75+*                                    16      0
      .****                                71      0
      .*                                   18      0
      .***                                 47      0
      .****                                58      |
      .*********                          147      |
      .**********                         376      |
      .**********                         621    +---+---+
      .**********                         706    *------*
      .**********                         820    +------+
      .**********                         417      |
-1.75+*                                    1      |
      ----+----+----+----+----+----+----+----+----+-
      * may represent up to 18 counts
```

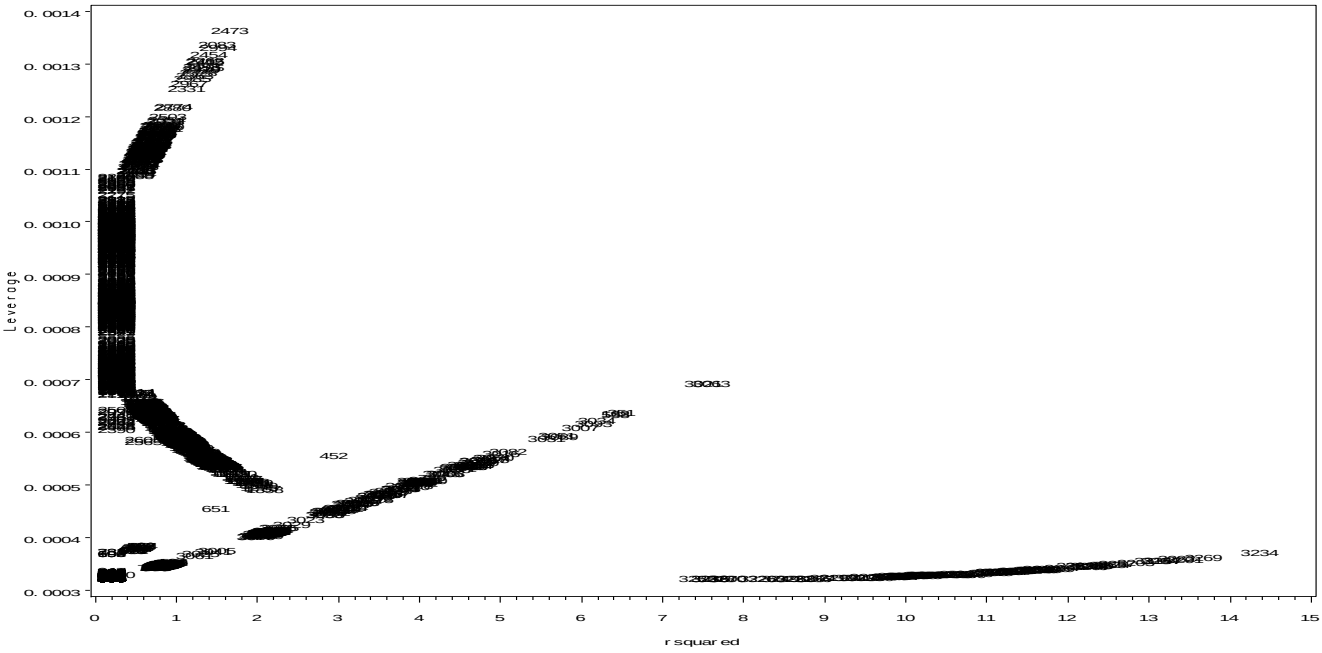
Second, we assess the leverages to identify observations that have a potentially large influence on regression coefficient estimates.

Figure 3: Leverage Distribution



After we closely examine the observations in the simulated dataset as plotted below, the “claim_pk” which is the ID number for claims in (3258,3236,3036,3270,3300,3228,3260,3136,3130,3111) displays high leverage. As a result, 200 claims with leverage>0.001 were identified as outliers.

Figure 4: "claim_pk" Plot for Leverage and R-squared



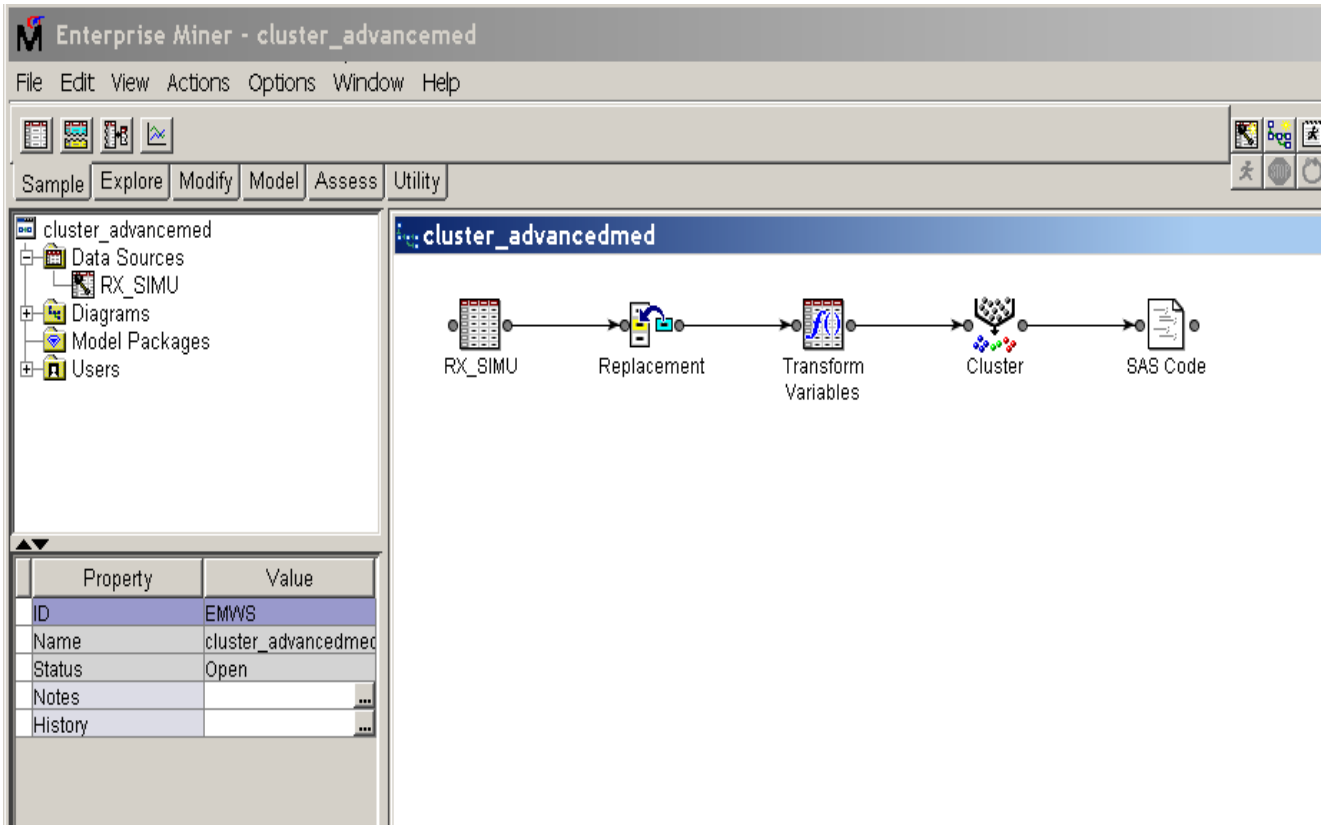
SAS code:

```
proc univariate data=rx_res plots;
  var lev;
run;
proc sql;
 create table rx_res2 as
 select *, r**2 as rsquared
 from rx_res;
quit;
goptions reset=all;
axis1 label=(r=0 a=90);
symbol1 pointlabel = ("#claim_pk") font=simplex value=none;
proc gplot data=rx_res2;
  plot lev*rsquared / vaxis=axis1;
run;
quit;
```

The results of Cook' distance showed that there were 118 claims with Cook's distance>4/3298 and 195 claims with an absolute value of DFFITS>2*sqrt(1/3298) which were considered as outliers.

We used the SAS(R) Enterprise Miner™ to do the cluster analysis. Each observation represents a claim for overpayment detection. The following is the flow diagram of this clustering model design.

Figure 5: Flow Diagram of the Clustering Model



In the "RX_SIMU" node, we did not use any target information created by a rule-based algorithm because it is not necessary for the unsupervised learning model. In the "Replacement" node, we replaced the missing value of character variables with "Unknown" and ignored the missing values of interval variables. In the "Transform Variables" node, we created a new variable "log_aq" by employing the formula: log_aq=log (AWP*quantity_of_service+1). To reduce the variance of the variable "AQ" which has a skewness of 17.76, a log transformation on "AQ" was performed and a new variable log_aq was created. Below are the statistics after the log transformation.

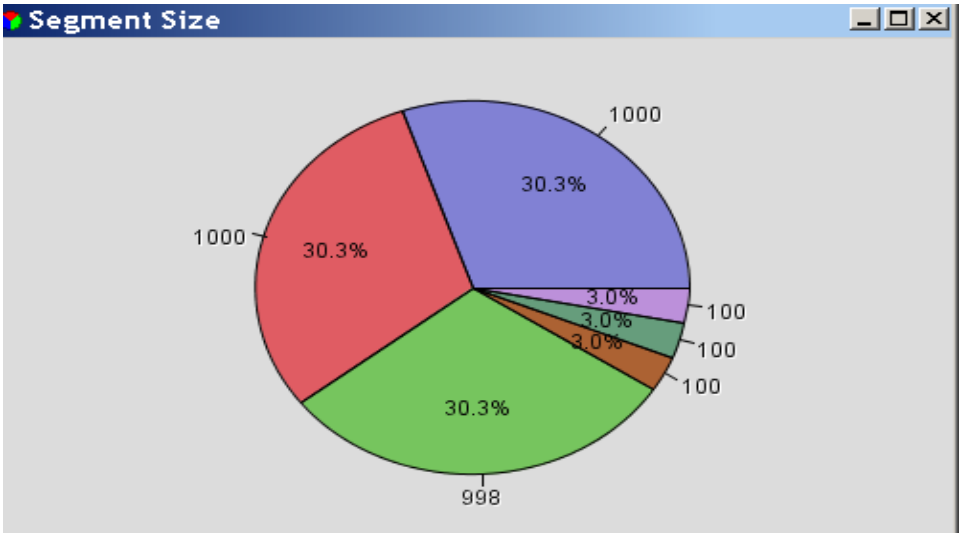
Figure 6: Transformation Statistics of "log_aq"

Transformations Statistics

Source	Type	Name	Formula	.	Missing	Minimum	Maximum	Mean	Std Devia...	Skewness
OUTPUT	Formula	log_AQ	log(awp_unit* quantity_of_service+1)	.	0	5.671397	8.81895	6.931477	0.65623	0.343216

The cluster selects initial seeds that are very well-separated using a full replacement algorithm. The following pie chart shows there are 6 segments selected for this clustering.

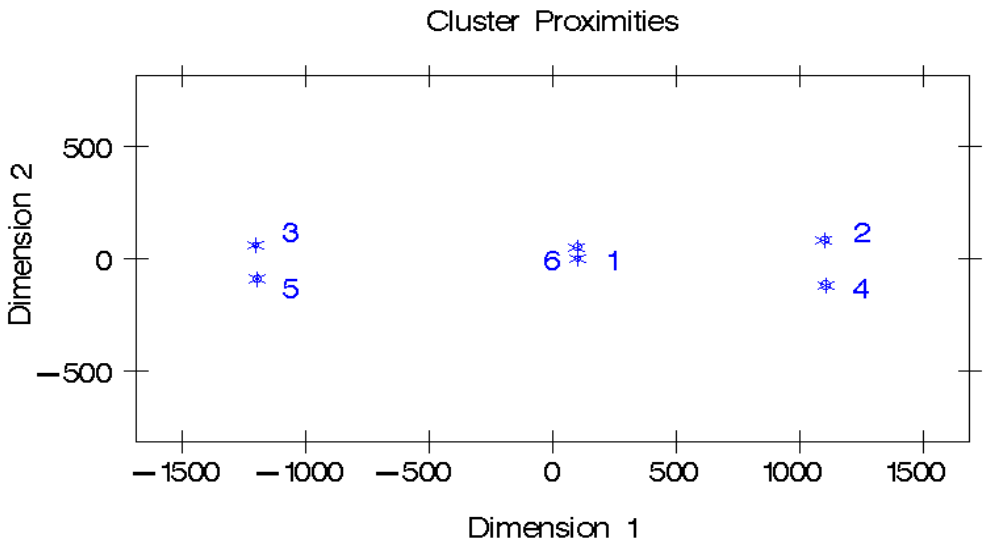
Figure 7: Segment Size Plot



There are 3 segments with sizes of around 1000 observations each, and 3 segments which have sizes of 100 observations each.

From the distribution of each variable within the segments, we know that most of them are evenly distributed within each segment and they appear the same in the pairs of (1, 6), (2, 4) and (3, 5).

Figure 8: Cluster Proximities Plot



Cluster proximity for average clustering is defined as the average pairwise distance of all pair of points from different clusters. From the plot of cluster proximities, the pattern becomes obvious. The distance of cluster proximities for the segment pairs of (1, 6), (2, 4) and (3, 5) are very close to each other. From the segment size plot, the sizes of segment 4, 5 and 6 are very small compared to their closest segments and hence can be identified as abnormal claims. After figuring out which variable caused this abnormality we used SAS code node to delete segments=4, 5 and 6. There are 300 claims in segments 4, 5 and 6 identified as abnormal claims.

SAS code:

```
libname cls  "C:\Documents and Settings\Administrator\Desktop\paper reference";
data cls.rx_clus;
  set &em_import_data.;
  if _segment_ in (1,2,3);
  drop _segment_ distance im_awp im_log_aq im_max_units
im_medicaid_amount_paid
        im_period im_quantity_of_service im_national_drug_code _impute_ log_aq;
run;
```

The following is the summary of experiment results for Student Residual, Leverage, Cook's distance, DFFITS and Clustering.

Table 2: Summary of Experiment Results

Statistical Techniques	Number of Abnormal Claims Removed	Abnormal Claims Capture Rate	Number of False Positives Removed	False Positives Capture Rate	Number of Normal Removed	Normal Claims Misclassification Rate
Student Residual	87	29%	0	0%	5	0.17%
Leverage	2	1%	0	0%	198	6.60%
Cook's Distance	195	65%	0	0%	3	0.10%
DFFITS	182	61%	6	67%	35	1.17%
Clustering	300	100%	9	100%	0	0.00%

CONCLUSION

When working with Medicaid data, AdvanceMed has learned that there are different types of data anomalies in Medicaid pharmacy claim data. A simulation of the pharmacy claim file shows that false positives are caused by these anomalies in a rule based algorithm. To avoid false positives, we introduced five different statistical approaches to detect and eliminate the abnormal claims. The results of this study indicate that clustering technique is the best approach, followed by DFFITS.

REFERENCES

[1]AdvanceMed Corporation
http://www.csc.com/advancemed/ds/11858-about_us

ACKNOWLEDGMENTS

Special Thanks to Tom Mathis, who is the program director of AdvanceMed Corporation, for his patience and support. Huge thanks to Rick Wells, who is the project director of AdvanceMed Corporation, for his incredibly understanding and sincere encouragement. Finally, to all of the colleagues who perfectly demonstrate creative excellences — thank you.

CONTACT INFORMATION

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

SHENJUN ZHU
Chief Statistician,
AdvanceMed Corporation,
2636 Elm Hill Pike, Suite 110, Nashville, TN 37214
p: 615.425.2481 | f: 615.872.0272
zhuc@admedcorp.com | www.admedcorp.com

QILING SHI
Data Analyst, Mathematics PhD
AdvanceMed Corporation,
2636 Elm Hill Pike, Suite 110, Nashville, TN 37214
p: 615.425.2451 | f: 615.872.0272
shiq@admedcorp.com, shiqiling@gmail.com| www.admedcorp.com

ARAN CANES
Data Analyst, Economics MA
AdvanceMed Corporation,
2636 Elm Hill Pike, Suite 110, Nashville, TN 37214
p: 615.872.0272 | f: 615.872.0272
canesa@admedcorp.com | www.admedcorp.com