Paper 1633-2014

Clustering and Predictive Modeling of Patient Discharge Records with SAS® Enterprise Miner™

Linda Schumacher, MS, Dr. Goutam Chakraborty
Oklahoma State University, Stillwater, OK

ABSTRACT

Can clustering improve a predictive model's overall fit statistic? This talk describes the methods and results of data mining pediatric IBD patient records. SAS® Enterprise Miner™ 12.1 was used to segment patients and model important predictors for the length of hospital stay using discharge records from the national Kid's Inpatient Database. Profiling revealed that patient segments were differentiated by primary diagnosis, operating room procedure indicator, comorbidities, and factors related to admission and disposition of patient. Cluster analysis of patient discharges improved the overall average square error of predictive models and distinguished predictors that were unique to patient segments.

INTRODUCTION

Healthcare information contained in medical claims and discharge reports has traditionally been analyzed by stringent statistical methods to calculate disease prevalence, utilization of resources, and factors influencing disease course. Data mining findings have the potential to complement those analyses by affirming and quantifying those findings or by discovering previously unanticipated relationships across healthcare data. The potential relationship factors discovered by analytic techniques such as predictive modeling, clustering and association analysis may offer new bases for future hypothesis-based analysis. Data mining of healthcare information can identify factors related to disease prevention and chronic illness management.

SAS® Enterprise Miner™ was used to identify the important predictors of the length of inpatient hospital stay for pediatric patients diagnosed with chronic Inflammatory Bowel Disease (IBD) using the Healthcare Cost and Utilization Project (HCUP) 2009 Kids' Inpatient Database (KID).

LITERATURE REVIEW

HCUP, which is sponsored by the Agency for Healthcare Research and Quality (AHRQ), provides a collection of databases containing samples of nationwide healthcare data. Review of the HCUP publication database in August 2013 found thirty papers whose abstracts included analysis of IBD patients. Of these, three were based on KID data (Nylund, 2013; Heaton, 2011; Pant, 2013). Nylund studied the risk of venous thrombotic events using logistic regression. Heaton et al. studied the 2006 hospital costs of pediatric IBD patients using ordinary least squares regression. Pant et al. studied the length of stays for 2009 IBD pediatric patients with a particular bacterial infection.

Another HCUP database is the National Inpatient Sample (NIS). Kaplan et al. (2013) used the HCUP NIS database and comorbidities software to study the effects of comorbidities on the outcomes of adult and pediatric IBD patients through logistic and linear regression models. Yang & Logan (2006) used data mining techniques to discover diseases related to paraesophageal hernia (PEH) in NIS data. The authors employed associate rule analysis to discover diseases that are associated with PEH. Each inpatient record was treated as a transaction with the multiple diagnoses as items in the transaction. The authors measured the support, or percent of transactions that contain both PEH and any other disease. Associate rule analysis results performed comparably with the findings of a survey of GI specialists.

A search of SAS conference proceedings found that data mining techniques are increasingly being applied in new ways to analyze healthcare data. Habek (2011) presented a methodology for disease management analytics. The method incorporated data preparation, segmentation, predictive modeling and tree maps. This method used clustering and segmentation of patients by medical and demographic information. Predictive models were developed for each cluster to predict hospital admission and diagnosis of thyroid disease. Habek's findings are easy to interpret and visualize because of the use of decision trees and tree maps. Cerrito & Cerrito (2006) analyzed hospital emergency department records. They performed association analysis and text mining of diagnoses and medications to determine differences in medical decision making. They also used transactional time series to forecast patient arrival time and length of emergency department visit.

DATA

HCUP integrates data from state level sources and includes information on all-payers, private and government based programs. In addition to the databases, related software tools and products are available for noncommercial use at the HCUP website http://www.hcup-us.ahrq.gov/. HCUP's Kids' Inpatient Database (KID) provides data on pediatric inpatient hospital stays for patients from birth to twenty years of age. HCUP provides de-identified hospital discharge records for research purposes only. The large sample size permits study of less prevalent diseases such as Inflammatory Bowel Disease (IBD). IBD patients' use of healthcare resources is detailed by Kappelman, et al. (2011). IBD diseases are Crohn's Disease (CD) or Ulcerative Colitis (UC).

The 2009 Kids' Inpatient Database consists of 3,407,146 records and provides a sample of discharges from 44 states. The KID does not detail information for individual patients. Each record consists of discharge level information. A single patient may have had multiple discharges but there is no way to identify this. All modeling, therefore, would relate to a hospitalization instance rather than an individual patient.

The core data set contained records with 159 variables covering information on diagnoses, procedures, patient demographics, and administrative data for each discharge. A hospital data set with 39 variables contained hospital demographic information. Forty comorbidity and disease severity variables were contained in the severity data set. An additional data set of Diagnosis and Procedure Groups (DRG) was available but was not used in this analysis. Hospital and record number variables were used only to merge records from multiple data sets.

Records for patients with a primary diagnosis of IBD were selected by examining the primary diagnosis variable, dx1. Discharges were selected by ICD-9 CM diagnosis code in the form the 555.xx for Crohn's Disease (CD) patients and 556.xx for Ulcerative Colitis (UC) (Kappelman p. 63). A primary diagnosis of IBD was present on 8499 records with an average length of stay of 6.658 days: 3171 indicated UC with an average length of stay of 6.475 days and 5328 indicated CD with an average length of stay of 5.679 days.

HCUP DATA USE COMPLIANCE

HCUP Data Use Agreement requires that no analysis report information potentially identifying patients or hospitals. Hospital and record number variables were not used as analysis variables but only used to merge records from multiple data sets. The smallest geographical unit variable used to specify hospital location is region.

To preserve anonymity, no stratified group of less than eleven records is permitted. Therefore the required minimum size of any cluster group or decision tree leaf must be at least twelve cases.

HEALTHCARE INFORMATION MODELING

Healthcare diagnoses and procedures are classified by coding schemes for a variety of medical and administrative agencies. The KID incorporates the following systems: ICD-9 CM2 and related Diagnosis and Procedure Groups (DRG), Clinical Classifications Software (CCS), Major Diagnostic Category (MDC), All Patient Refined DRGs (APR-DRGs) and Disease Staging V5.2 Diagnostic Categories (DS). These different systems provide valuable but often overlapping or correlated information.

This analysis selected variables that were either directly-coded or derived from ICD-9 encoded variables. HCUP comorbidity software used ICD-9 and DRG codes to create comorbidity indicators. The variable names for these indicators are prefaced with the characters CM_. The DS: Stage of Principal Disease Category, DS_stage1, variable measured the severity of the primary diagnosis. The Disease Staging Categories, GIS37 and GIS09, correspond to UC and CD ICD-9 codes. Therefore, the variables DX1, CM_condition, and DS_Stage1 are compatibly used and this analysis limited the use of diagnosis information to DX1, the primary diagnosis, and the twenty nine HCUP comorbidity indicators.

VARIABLES

Variables in the KID database containing medical, administrative and demographic information were selected for modeling.

Variable Name	Measurement Level	Label
DX1	nominal	ICD 9 –CM Primary Diagnosis
DS_Stage1	interval	Disease Staging: Principal Stage
LOS	interval	Length of stay (HCUP cleaned)
NCHRONIC	interval	Number of chronic conditions

NDX	interval	Number of diagnoses on this record
NPR	interval	Number of procedures on this record
DISEASE	binary	Disease Indicator 0-UC 1-CD
ELECTIVE	binary	Elective versus non-elective admission
ORPROC	binary	Major operating room procedure indicator
DISPUNIFORM	nominal	Disposition of patient (uniform)
HCUP_ED	binary	HCUP Emergency Department service indicator
TRAN_IN	nominal	Transfer in indicator

Table 1. HCUP KID Medical and Administrative Variables

Variable Name	**Measurement Level**	**Label**
AGE	interval	Age in years at admission
AWEEKEND	binary	Admission day is a weekend
DQTR	ordinal	Discharge quarter
FEMALE	binary	Indicator of sex 0-Male 1-Female
PAY1	nominal	Primary expected payer (uniform)
PAY2	nominal	Secondary expected payer (uniform)
RACE	nominal	Race (uniform)
ZIPINC_QRTL	ordinal	Median household income
HOSP_CONTROL	nominal	Control/ownership of hospital
HOSP_LOCATION	binary	Location (urban/rural) of hospital
HOSP_REGION	nominal	Region of hospital
HOSP_TEACH	binary	Teaching status of hospital
HOSP_BEDSIZE	ordinal	Bed size of hospital

Table 2. Patient and Hospital Demographic Variables

DATA PREPARATION

HCUP merged data from the separate state databases and integrated variables with the same information but different coding schemes into uniformly coded variables. The labels in Tables 2 and 3 indicated which variable have been recoded and/or cleaned by HCUP.

Variables for admission source (ASOURCE) and Secondary expected payer (PAY2) had over 75% missing and were rejected.

The distributions of medical related variables are often not normal being left or right skewed. Variables for number of procedures, length of hospital stay, number of chronic illnesses were right skewed in the 2009 KID database. The age variable was left skewed. To facilitate cluster and regression analysis, skewed variables were transformed to variables that follow more normal distributions using the Transformation node. The right skewed variables were transformed using `log(var + 1)`. Left skewed age was transformed using the maximize normality method. The formula generated by the transform node was `(max(AGE-0, 0.0)/20)**4)`.

PREDICTIVE MODELING

Enterprise Miner offers many types of nodes that perform predictive modeling for interval targets. Because the primary objective of this analysis is inference, the regression and decision tree nodes were used. These models facilitate finding and interpreting important predictors. Decision trees were built with the two splitting rules criteria available for interval targets and sub-branches were optimized using Average Square Error statistic. Linear regression models were developed using the stepwise variable selection method. Both main effects and interaction effects models were created and tested.

Assessing the models was completed using the Partition Node and Model comparison node. The Partition Node separates the data into train and validation sets. The regression and decision trees nodes develop the model on the training data and provide output results for both training and validation sets. The model comparison node compares all input models and determines the model with the best validation selection statistic. By selecting the best validation statistic the chosen model is less likely to be over-fit to the training data. The model with the best validation Average Square Error value of 27.35 squared days was a decision tree built with using the ProbF, probability of F statistic, splitting rule for an interval target.

Preliminary predictive modeling found that the top two most important variables for determining length of stay were number of diagnosis and procedures, an expected result. Segmenting the discharge records into groups with similar characteristics and performing predictive analysis on each segment separately is likely to yield better results.

CLUSTERING

To improve the modeling, hospital discharge records were segmented by five dominant medical predictors using clustering. Clustering was first performed in SAS® Enterprise Miner™ with the clustering node setting using range standardization of variables and centroid clustering method for initial stage followed by k-means in the second stage. Initial cluster results were unacceptable having single observations joining in the final clustering rounds. The data was then examined in SAS$^{®}$ Enterprise Guide™. PROC CLUSTER with trimming option of 5% was used to reduce distortion of clusters by outliers. After trimming, there were 8,069 records UC: 2972, and CD: 5097. This trimmed data was then used as input data to SAS® Enterprise Miner™ clustering node with 6 clusters k-means algorithm using numbers of procedures, diagnoses, and chronic illness and indicators for disease type and major operating room procedure.

Segment Number	Segment Size: n_i	Percentage
1	4274	53.0
5	1437	17.8
4	1098	13.6
6	437	5.4
3	415	5.1
2	408	5.1

Table 3. Segment Sizes

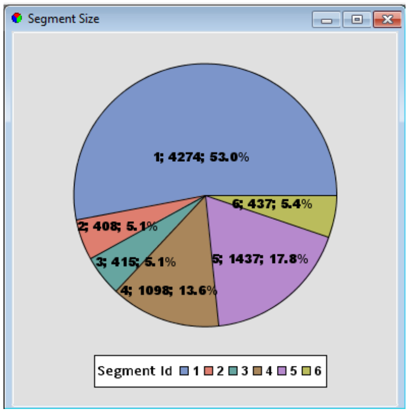


Figure 1. Segment Size Pie Chart

SEGMENT PROFILING

The Segment Profile Node reports on variables whose distributions within a cluster are substantially different from the distribution of the same variable in the overall data. The values of these variables create a profile of the cluster. The minimum worth parameter sets the minimum worth threshold for a variable to be included in the profile. The six segments were profiled separately by the base variables, those used to cluster the variables, and the descriptor or demographic variables those not directly used in clustering. The minimum worth parameter was set to 0.001 for the base variables and 0.003 for the descriptors. Patient segments were differentiated by primary diagnosis, operating room procedure indicator, comorbidities, and factors related to admission and disposition of patient.

Figure 2 shows a portion of the profile node's output for cluster 2. The red bar outline in the histograms shows the distribution of an interval variable the entire data set. The blue solid bars show the distribution of the variable in segment 2. Binary and nominal variable proportions are shown in concentric pie charts. The center circle shows proportion in the overall data set. The outer ring displays the proportion in cluster 2.

The profile of cluster 2 indicates that the patients had Crohn's Disease, a major OR procedure, were more likely to have a higher number of procedures than overall, were more likely lower number of chronic diagnosis, and had a higher proportion of elective admissions than overall. Table 4 displays the profiles for each segment.

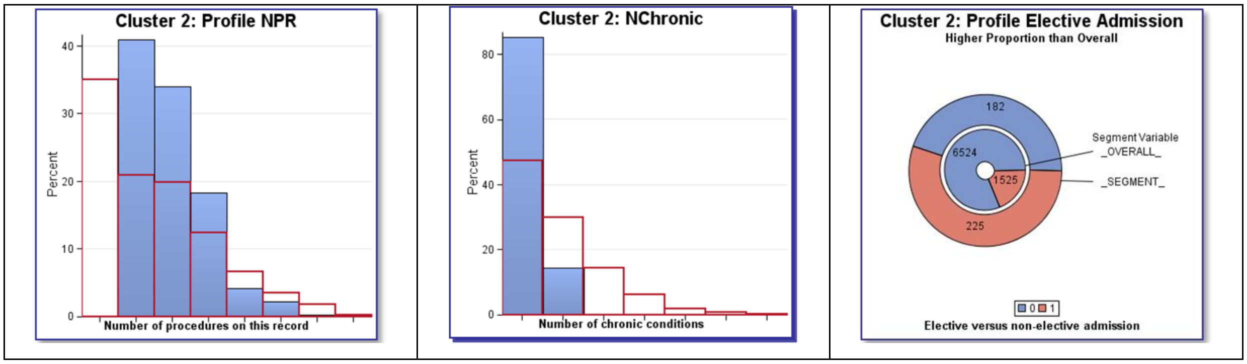


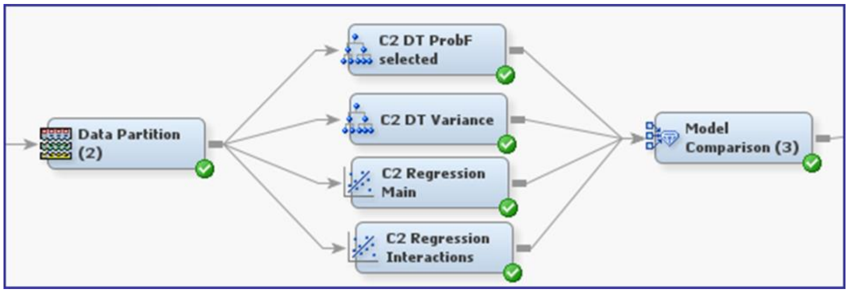
Figure 2. Profile Plot for Cluster 2

Segment Id	dx1	Major OR	More Likely Lower Distribution or Proportion Than Overall	More Likely Higher Distribution or Proportion Than Overall
1	CD	no	# procedures, Elective Admissions	Emergency Dept. service
2	CD	yes	# Chronic Conditions	# procedures, Elective Admissions
3	CD	yes		# procedures, # diagnoses # Chronic Conditions
4	UC	no		# diagnoses, # Chronic conditions, Anemia CM, Electrolytes CM
5	UC	no	# procedures, # diagnoses # Chronic Conditions	
6	UC	yes	Emergency Dept. service	# procedures, Elective Admission, Discharged to Home Care

Table 4. Profiles by Segment

PREDICTIVE MODELING BY CLUSTER

Data for each cluster was partitioned into equal parts for training and validation. Predictive modeling using decision tree nodes and linear regression nodes were developed separately for each cluster. The best predictive model for each segment was chosen by best validation Average Square Error (ASE). For each cluster the model with the lowest ASE was one of the decision tree models.



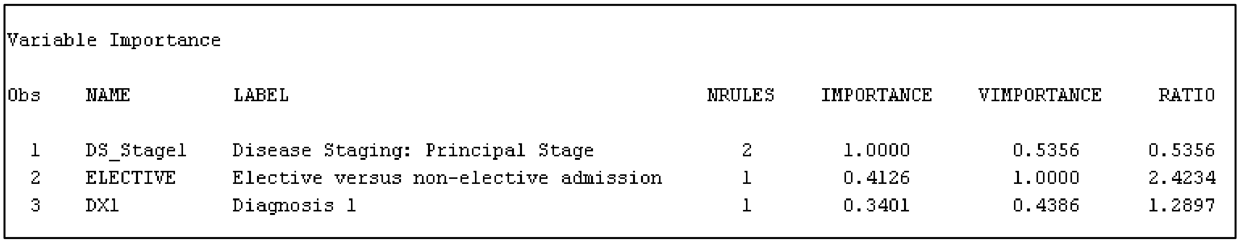
Display 1. Process Flow Diagram

Cluster	Splitting Criterion	ASE
1	ProbF	11.29
2	ProbF	7.89
3	ProbF	39.05
4	ProbF	29.78
5	ProbF	10.11
6	Variance	42.44

Table 5. Smallest ASE by Cluster

Skewed variable transformation is not a requirement for building decision trees. The tree models were therefore recreated using the original non-transformed variable.

Output of the decision tree model includes a listing of the most important variables by variable importance statistic, a measure of relative importance. The Variable Importance for segment 2 is found in Output 1.



```
Variable Importance

Obs    NAME         LABEL                                    NRULES    IMPORTANCE    VIMPORTANCE     RATIO

  1    DS_Stage1    Disease Staging: Principal Stage            2        1.0000         0.5356      0.5356
  2    ELECTIVE     Elective versus non-elective admission      1        0.4126         1.0000      2.4234
  3    DX1          Diagnosis 1                                 1        0.3401         0.4386      1.2897
```

Output 1. Results Segment 2 Variable Importance

The profile information and important predictors for each segment are found in Table 6. Predictors are listed in decreasing order of importance. The number of procedures and diagnoses, the elective admission indicator, and the stage of the primary disease were verified as important predictors. Predictors unique to individual clusters were determined to be anemia and electrolyte comorbidities, a particular primary diagnosis, household income, race, bed size, region, teaching status of hospital, and disposition of the patient.

dx1	Major OR	Profile	Important Predictors	Segment Id
CD	no	Likely to have fewer procedures, Higher proportion of non-elective admissions, and Emergency Dept. services	NPR, NDX, Disposition of patient	1
	yes	Likely to have fewer chronic conditions and more procedures and elective admissions	Disease stage, Elective admission, Diagnosis (dx1)	2
		Likely to have more procedures, diagnosis, and chronic conditions	NPR, Elective admission, Teaching hospital	3
UC	no	Likely to have more procedures, chronic conditions and a higher proportion of Blood Loss Anemia CM and Electrolytes CM	NPR, NDX, Deficiency Anemia, Weekend admission	4
		Likely to have fewer procedures, diagnoses, and chronic conditions	NPR, Median HH Income	5
	yes	Lower proportion of Emergency Dept. services, Higher number of procedures, elective admissions, discharged to home care	Elective admission, Disease stage, NPR, Race, Bed size, Region	6

Table 6. Profile Variables and Important Predictors by Segment

DISCUSSION

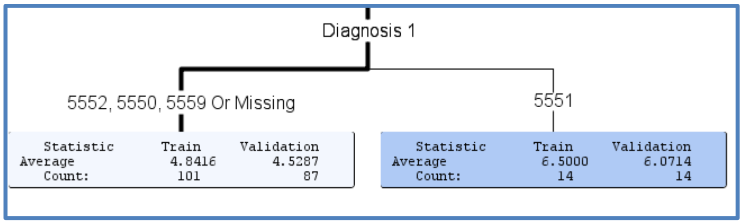
Did clustering improve the overall fit statistic? The initial predictive model ASE was 27.34 square days. In the process of clustering the more extreme outliers were trimmed from the data. The predictive model was refit with the trimmed data and resulted in a validation ASE of 17.90 square days. An overall weighted Average Square Error for the clustered model is calculated by computing a weighted sum of the individual cluster ASE statistics.

$$overall\ ASE = \sum_{i=1}^{6} \frac{n_i}{N} * ASE_i$$

where n_i is the number of observations in segment *i*, *N* is the total number of observations, and ASE_i is the Average Square Error for segment *i*. The resulting overall ASE for the segmented model was 16.53 square days. Cluster analysis of patient discharges improved the overall average square error of predictive models by 7.7%. Future modeling using both the primary and all other diagnoses on the discharge record from clustering would likely show an even larger improvement.

Profiling revealed that patient segments were differentiated by primary diagnosis, operating room procedure indicator, comorbidities, and factors related to admission and disposition of patient Table 6.

Generally patients with less chronic conditions, number of procedures and diagnoses, were predicted to have lower lengths of stay. Predictors that were unique to patient segments 2 and 4 may provide new insights. Display 2 shows a sub-tree branch for the segment 2 model. Patients with a particular ICD-9 code as the primary diagnosis were predicted to have a higher length of stay than any other diagnosis code.



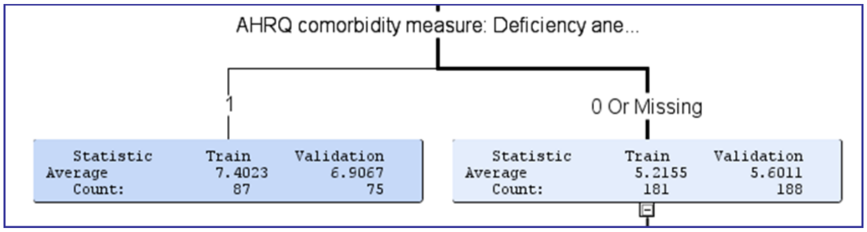
Display 2. Segment 2 Tree Branch: Stage < 2.06 and elective admission

Cluster 2 discharges had Crohn's Disease, a major OR procedure, were more likely to have a higher number of procedures, were more likely to have a lower number of chronic diagnosis, and had a higher proportion of elective admissions. The important predictors were disease stage, elective admission, the ICD- 9 which indicates the location of the Crohn's disease. Patients with elective admissions were predicted to have lower lengths of stay than those with non-elective admissions. Holding elective admission constant, patients with less severe staging were predicted to have lower lengths of stay. For patients with elective admission and disease staging of <= 2.05, the ICD-9 code in dx1 was an important predictor. Those patients with code 555.1, which indicates enteritis of the large intestine were predicted to have lower lengths of stay than patients with disease in the small intestine, both large and small intestine, or not specified.

Elective Admission	**no**		**yes**			
Disease staging	**<=2.02**	**2.03-2.04**	**<= 2.05**			**>= 2.06**
			dx	**5550,5552,5559**	**5551**	
LOS	**5.0769**	**7.7308**	**LOS**	**4.8416**	**6.5000**	**8.375**

Table 7. Cluster 2 Predicted Lengths of Stay

Cluster 4 discharges had Ulcerative Colitis, did not have a major operating room procedure, were more likely to have a higher number of diagnoses and chronic conditions, and had a higher proportion of blood loss anemia and electrolytes comorbidities. The important predictors were the number of procedures and diagnoses, the deficiency anemia comorbidity and weekend admission indicators. Display 3 contains the sub-tree branch for discharges with 1 or 2 procedures. Those patients with a deficiency anemia were predicted to have a higher length of stay, 7.40 days, versus the stays of those without, 5.2 days.


AHRQ comorbidity measure: Deficiency ane...
1
0 Or Missing
Statistic Train Validation
Average 7.4023 6.9067
Count: 87 75
Statistic Train Validation
Average 5.2155 5.6011
Count: 181 188

Display 3. Cluster 4 Tree Branch 1-2 Procedures

These predictors unique to particular segments may warrant further investigation by hypothesis based controlled studies.

CONCLUSION

Hospital discharge records were analyzed to identify important predictors for the length of hospital stays for pediatric IBD patients. Initial predictive models were unsatisfactory. A method of segmenting the original data into similar groups of patients and performing predictive modeling by segment improved the overall fit statistic. Patient segments were differentiated by primary diagnosis, operating room procedure indicator, comorbidities, and factors related to admission and disposition of patient. The number of procedures and diagnoses, the elective admission indicator, and the stage of the primary disease emerged as important predictors. Longer hospital stays are predicted for patients with higher number of procedures, diagnoses, and chronic conditions as expected. Predictors unique to individual clusters included an anemia comorbidity, a particular primary diagnosis, teaching status of hospital, and disposition of the patient. A subset of Crohn's patients with a diagnosis of disease located only in the large intestine and a subset of ulcerative colitis patients with anemia where predicted to have longer hospital stays.

This analysis explored using data mining techniques to infer important predictors of length of hospital stay using one data set. Further studies on new data sets should examine if the patterns found for this particular data are replicated. The authors do not have medical domain expertise. This information is therefore not intended nor otherwise implied to be medical advice of hospital stay for patients.

This analysis also explored using segmented versus global models for inference. Clustering the observations into segments and creating predictive models by segment improved the assessment statistic, ASE, by 7.7%. Further research should be performed to replicate this method on other data sets to gather evidence that segmented models can outperform global models. Additional simulation studies should be performed to determine under which boundary conditions this result could be generalized.

REFERENCES

Cerrito P. & Cerrito J. (2006). Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs. Proceedings of the Thirty-first Annual SAS® Users Group International Conference, Paper 077-31

Habek, G. S. (2011). Taking Disease and Health Management Analytics into the Next Generation. Proceedings of the SAS® Global Forum 2011 Conference, Paper 182-2011.

HCUP Nationwide Inpatient Sample (NIS). Healthcare Cost and Utilization Project (HCUP). 2007-2009. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/nisoverview.jsp

HCUP Comorbidity Software. Healthcare Cost and Utilization Project (HCUP). 2009. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp. Accessed August 2013.

HCUP Kids' Inpatient Database (KID). Healthcare Cost and Utilization Project (HCUP). 2006 and 2009. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/kidoverview.jsp

Heaton, P. C., Tundia, N. L., Schmidt, N., Wigle, P. R., & Kelton, C. M. (April 2012). National Burden of Pediatric Hospitalizations for Inflammatory Bowel Disease: Results from the 2006 Kids' Inpatient Database. Journal of Pediatric Gastroenterology and Nutrition, ISSN 0277-2116, (54, 4), 477–485.

Kaplan,G. G., Hubbard,J., Panaccione,R., Shaheen,A. A., Quan,H., Nguyen,G. C., … Myers,R. P. (Aug 2011). Risk of Comorbidities on Postoperative Outcomes in Patients with Inflammatory Bowel Disease. Archives of Surgery, (146, 8),959-964.

Kappelman, M.D., Porter, C.Q., Galanko, J.A, Rifas-Shiman, S.L., Ollendorf, D.A., Sandler, R.S., & Finkelstein, J.A. (2008). Direct Health Care Costs of Crohn's Disease and Ulcerative Colitis in United States Children and Adults. Gastroenterology, (135) 1907–1913.

Nylund,C. M., Goudie, A., Garza, J. M., Crouch, G., & Denson, L. A. (May 2013) . Venous Thrombotic Events in Hospitalized Children and Adolescents with Inflammatory Bowel Disease. Journal of Pediatric Gastroenterology and Nutrition, (56, 5), 485-491.

Pant, C., Anderson, M. P., Deshpande, A., Altaf, M. A., Grunow, J. E., Atreja, A.,& Sferra, T. J. (April 2013), Health Care Burden of Clostridium Difficile Infection in Hospitalized Children with Inflammatory Bowel Disease, Inflammatory Bowel Diseases, (19,5), 1080-1085.

Yang,J.,& Logan,J. A Data Mining and Survey Study on Diseases Associated with Paraesophageal Hernia.(2006) AMIA Annual Symposium Proceedings, 829-833.

ACKNOWLEDGMENTS

The authors thank Michael Kappelman, MD, MHS, an associated professor at UNC Children's Hospital, for his guidance in selecting HCUP pediatric IBD patient data and the length of hospital stay target variable. The authors are grateful to HCUP for providing nationwide databases to student researchers at a reduced rate.

CONTACT INFORMATION

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

Dr. Goutam Chakraborty
Oklahoma State University
Stillwater, OK, 74078
goutam.chakraborty@ okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

Linda Schumacher
8517 Bluebill Ct.
Raleigh, NC 27615
919-848-1374
schumachers@bellsouth.net

Linda Schumacher received her Graduate Certificate in Business Data Mining from Oklahoma State University in December 2013. She holds a Master of Science with thesis in computer science from the University of Illinois at Urbana Champaign. She earned her Bachelor of Science with high honor in applied mathematics/computer science from Stevens Institute of Technology in Hoboken, NJ. She has 10 years of experience in database programming and software engineering. She is a SAS® Certified Advanced Programmer for SAS® 9, Certified Predictive Modeler Using SAS® Enterprise Miner™ 7, and has received her SAS® and OSU Data Mining Certificate. Her OSU team placed 3rd in the Analytics 2013 SAS® data shootout.