EXPLORATION OF THE OUTCOME OF PATIENT DEMOGRAPHICS, BEHAVIOR AND INTENTIONS ON COST OF CARE

Abstract

In *Predictive Analytics and the New World of Retail Healthcare,* Simmi Singh and Tia Sawhney describe the introduction of predictive analytics to healthcare by health insurers. Healthcare has lagged behind other industries in using predictive analytics to innovate and change user experience for a number of reasons including availability of sufficient "statistically valid" data. This study was conducted using data extracted from the athenahealth cross-practice data warehouse in Watertown, MA. The data incorporates patient data from approximately 5,000 healthcare practices of varying size across 6 states in the United States: MA, CA, FL, TX, OH, and NY. De-identified patient data was sampled from 625,190 unique patients for 2014. The R packages randomForest, caret, adabag, rpart, ada, C50, tree and party were used to perform C50 decision tree and random forest bagging ensemble algorithms on sample subsets of the 625,190 patients. The decision tree model accurately predicted 121,683 patients annual cost category out of 156,298 patients in the testing dataset (79%). A 79% accuracy, although modest, instruments the potential for increased patient health through decision tree and random forest learning algorithms as applied to patient data. These techniques will be increasingly important in the healthcare landscape as ACOs look to succeed where HMOs failed.

Introduction

In *Predictive Analytics and the New World of Retail Healthcare,* Simmi Singh and Tia Sawhney describe the introduction of predictive analytics to healthcare by health insurers. Healthcare has lagged behind other industries in

using predictive analytics to innovate and change user experience for a number of reasons including availability of sufficient "statistically valid" data. Aggregating statistically valid data has been so hard due to the segmentation of patient's health information across the continuum of care. Health insurers are just touching the surface of the power of predictive analytics and machine learning being applied to patient data.[i]

Today, healthcare insurers are starting to launch products using psycho-demographics as correlated to healthcare utilization. Also, products such as Symmetry and DxCG are leading the healthcare analytics market with products using machine learning to forecast future claim volumes. These products are using detailed health demographic information. However, due to the level of detail these products require for accuracy, they are mainly used only at renewal of payer contracts. As healthcare data gets mined more and more, predictive analytics presents the only true promise in matching healthcare costs with actual healthcare consumers.[ii]

Additionally, machine learning techniques are being applied to longitudinal patient claims data to predict costs within chronic conditions consumers. In *Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: data-driven predictive analytics in healthcare,* Jon Maguire and Vasant Dhar from NYU's Stern School of Business use decision trees to make some interesting discoveries within the chronic diabetes population in the United States. Expectedly, they confirmed that patients in older age groups have higher medical charges. Also, patients with less advanced disease regimens were expected to have less costly healthcare utilization during the observation period. Less expectedly, they observed that patients in the Midwest and South regions would be expected to have cheaper medical charges compared to patients in the West and Northeast. Finally, patient's specifically using biguanide+DPP4 are associated with less than expected medical expenses but few people in the study were using this treatment. In this way, the analysis begs the question of why aren't more type 2 diabetics using this treatment regimen. Such predictive analyses, using longitudinal patient data, highlights the ability of machine learning and predictive analytics to further the science of evidence-based medicine.[iii]

Another promising application of machine learning within the healthcare space is the development of Accountable Care Organizations (ACOs). ACOs stand at the crossroads of predictive analytics in healthcare having both the financial incentive and providers under one roof to rally around the banner of population health. One article accurately states the good news that hospitals and health systems are already storing and collecting an inanimate amount of patient, clinical and financial data. However, these health systems do not have the resources or wherewithal yet to scale their technology to machine learning via data warehouses.[iv] Additionally, it is important to incorporate missing data elements into predictive models, especially in healthcare data.[v] Even with the ACO push in full fledge, patient data is never fully complete or fully defined. As

modeling expands in the age of population health, machine learning techniques will play a huge role in defining the most important indicators of patient health.

Methods

This study was conducted using data extracted from the athenahealth cross-practice data warehouse in Watertown, MA. The data incorporates patient data from approximately 5,000 healthcare practices of varying size across 6 states in the United States: MA, CA, FL, TX, OH, and NY. De-identified patient data was sampled from 625,190 unique patients for 2014. In order to make it into the sample set, patients had to still be living through 2014 and had to have encountered a practice using athenahealth's clinical technology services.

Of the 625,190 patients, 15 fields were identified from data captured in athenahealth's database through the cloud: state, gender, marital status, ethnicity, primary language, race, access to athenahealth's patient portal, access to a mobile phone, patient's consent for provider to call, age category, primary insurance, chronic conditions flag, patient medications entered online, claim volume category, and cost category. All data out of the 15 variables were associated with the annual year 2013 other than cost category which was associated with annualized charge volume per patient for 2014. Of these 15 flags, 5 fields were yes/no flags. Access to the patient portal, consent to call, access to mobile, chronic conditions, and medications entered online were all yes/no flags predicated upon the patients consent or availability for each of these fields. Primary Language and Ethnicity were cleaned to include only the top 3 with the rest being defined as other. Primary Language included English, Patient Declined, and Spanish while ethnicity included Not Hispanic or Latino, Hispanic or Latino and Patient Declined. Race was categorized by the top 6 race entries: White, Patient Declined, African American, Other Race, Asian, American Indian or Alaska Native, and Other Race. All NA fields were considered Missing fields and were not discluded from the dataset.

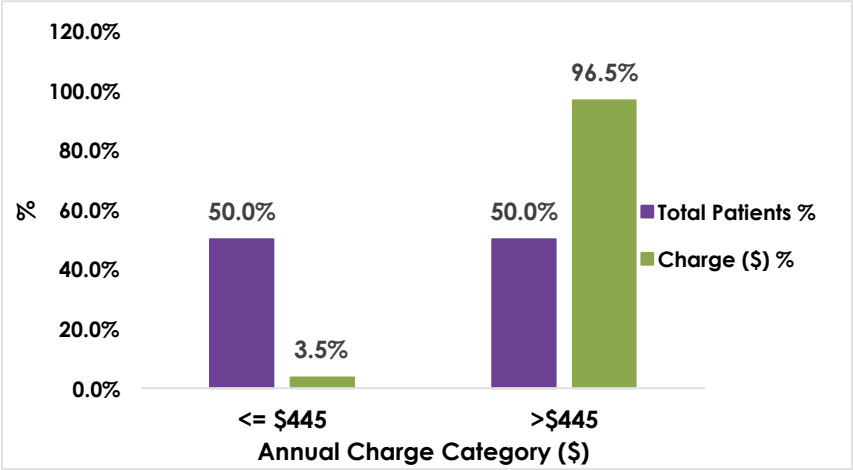
Additionally, the total claims variable included the following claim categories for 2013: <=1, 2,3,4,5,>5. Also, the cost variable was based upon 2014 annual charges associated with unique patients. This charge volume figure was then categorized into halves based on patient volume. The first half was less than or equal to $445.00 annual charge volume per patient. The point of the study was to explore whether or not these fourteen 2013 indicator variables could accurately predict whether or not patients fell into certain cost groupings in 2014.

R programming was used to process and clean the data to meet these categorical criterion and remove outliers. There were around 100 outliers which were un-trustworthy data points that associated patient's negative annual costs or abnormally high charge volume greater than $1 million. The R packages

randomForest, caret, adabag, rpart, ada, C50, tree and party were used to perform C50 decision tree and random forest bagging ensemble algorithms on sample subsets of the 625,190 patients.

Results

The 625,190 patients sampled from athenaNet practices show that 97% of the costs in these healthcare settings were born by the top 50% of the patients from a charge standpoint. In this study, the median annual cost per patient was $445.00 dollars in 2014. As a matter of statistics, the half of the patient population above $445 dollars accounts for most of the healthcare costs in fiscal year 2014.



Firgure 1: Top half of patients represent 97% of helathcare cost in 2014

Using the C5 algorithm to complete a decision tree analysis, Figure 2 illustrates the cross-validation results of the output in R. The best model was chosen as 2 nodes with an X-val Relative Error of around 0.56.

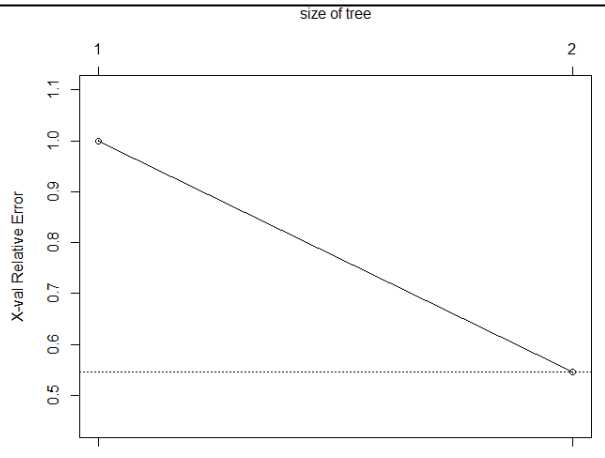


Figure 2: Cross-validation table of the C5 decision tree algorithm as applied to the 468,892 test patients.

Table 1 reveals the primary and surrogate splits in the nodes which illustrate some interesting discrepancies. Total_Claims_Flag or total claims in fiscal year 2013 is by far the most important indicator of total patient cost in 2014. In the table, left son corresponds to <=$445 while the right son corresponds to >$445. Therefore, in the primary split, the only claim category that is being mapped to the left son is <=1 claim. In other words, as claim volume increases per patient in 2013, cost increases in 2014. Age is the next most important indicator. In agreement with most literature to date, the two most costly age groups that are mapping to the right son are the two oldest groups: >54 and <=68 and >68. Interestingly, the missing age category is mapping to the less costly grouping.

In terms of primary insurance, the Insurance Reporting Category grou ps that are mapping to th

actual cost category	predicted cost category <=$445	>$445	Row Total
<=$445	59244	16377	75621
	0.379	0.105	

Table 2: Cross-tabulation of predicted versus actual cost categories of the C5 decision tree algorithm applied to the testing patient data and predicted on training data set. The training data set includes 156,298 patients sampled

e right son or most costly group are Legal/MVA, Medicare A, Medicare B, and W

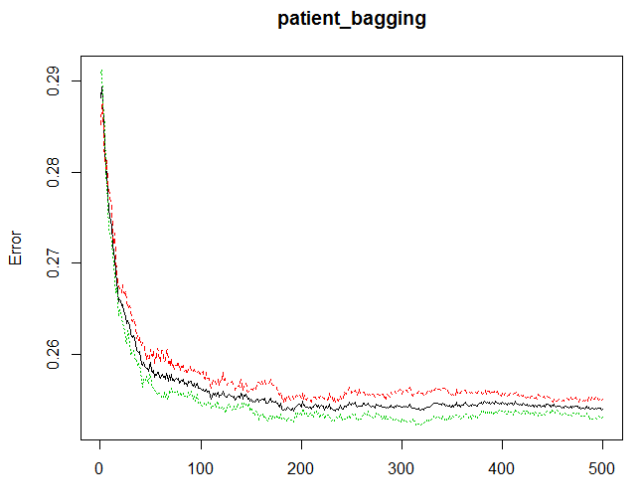
```
Node number 1: 468892 observations,    complexity param=0.4549639
  predicted class=>$445   expected loss=0.4828084  P(node) =1
    class counts: 226385 242507
   probabilities: 0.483 0.517
  left son=2 (237675 obs) right son=3 (231217 obs)
  Primary splits:
      Total_Claims_Flag splits as  RLRRRR,      improve=52723.850, (0 missing)
      Age               splits as  LLRRL,       improve= 2745.081, (0 missing)
      Primary_Insurance splits as  LLLLRLRRLLR, improve= 2650.598, (0 missing)
      Portal_Access     splits as  LRR,         improve= 1748.364, (0 missing)
      Marital           splits as  -RRRRLRRR,   improve= 1453.175, (0 missing)
  Surrogate splits:
      ConsentToCall     splits as  LRR,         agree=0.556, adj=0.099, (0 split)
      MedsEnteredOnline splits as  LRR,         agree=0.556, adj=0.099, (0 split)
      Primary_Insurance splits as  LLLRRRLRLLR, agree=0.532, adj=0.050, (0 split)
      Age               splits as  LLLRL,       agree=0.528, adj=0.044, (0 split)
      State             splits as  RLRLLR,      agree=0.524, adj=0.035, (0 split)

Node number 2: 237675 observations
  predicted class=<=$445  expected loss=0.2833239  P(node) =0.5068864
    class counts: 170336 67339
   probabilities: 0.717 0.283

Node number 3: 231217 observations
  predicted class=>$445   expected loss=0.2424086  P(node) =0.4931136
    class counts: 56049 175168
   probabilities: 0.242 0.758
```

Table 1: Summary of the fit of the C5 algorithm as applied to 625,190 patients across MA, FL, TX, CA, OH, and NY. The largest predictor of patient cost in fiscal year 2014 was previous claim volume in 2013.

orkers Compensation. Intriguingly, self-pay and missing insurance groups are ass ociated with the less expensive group. Also, portal access shows that the most e xpensive patients are correlated to having an answer either yes or no to the port al access question. Missing data in this field is associated with the least expensiv e patient cohort. Finally, marital status indicates that only single patients are bein g grouped into the least expensive patient grouping.

Table 2 displays a predicted versus actual cost category table of the C5 algorithm applied to the training patient dataset. Out of the 156,298 patients in the training dataset, 121,683 patients were mapped to the correct cost category. This table exemplifies an accuracy of 77.9%.

For fear of overestimating the prediction accuracy of the decision tree algorithm applied to this specific training dataset, the bagging random forest technique was used to select the mean result of 500 tree iterations. Figure 3 exemplify's the converage of error decreasing over random tree iterations with a



leveling off at around the 50th iteration to the 500th iteration.

Random forest bagging resulted in a slightly decreased prediction accuracy on the testing dataset of 74.7%. However, the bagging technique shows agreement on the important of Indicators as can be shown by the Mean Decrease Gini numbers in table 3. As in the C5 decision tree model, Claim Volume, Age, Primary Insurance, and Marital again seem to be the most valuable indicators for predictive insight into cost.

Figure 3: Error convergence across 500 tree iterations of random forest bagging.

Indicators	<=$445	>$445	MeanDecreaseAccuracy	MeanDecreaseGini
State	0.040	0.048	0.044	2266.674
Gender	0.003	0.006	0.005	1320.303
Marital	0.022	0.023	0.022	2371.018
Ethnicity	0.020	0.027	0.024	1410.236
Primary Language	0.005	0.035	0.021	939.051
Race	0.025	0.035	0.030	1537.035
Portal Access	0.033	0.045	0.039	1065.387
Mobile Phone	0.017	0.014	0.015	1710.351
Consent to Call	0.006	0.013	0.009	749.508
Age	0.041	0.020	0.030	1852.787
Primary Insurance	0.036	0.010	0.023	2389.898
Chronic Conditions	0.000	0.001	0.000	83.434
Medications Entered Online	0.006	0.014	0.010	754.579
Claim Volume	0.134	0.144	0.139	8437.899

Table 3: Important table of random forest bagging model on the training patient dataset. Shows the mean Gini decrease which indicates the entropy or information gain of each indicator.

Discussion

In distinction from the Dhar/Maguire study, this studies results highlight the predictive value of the combination of clinical, financial and demographic data in the healthcare space. In agreement with the decision tree created by Dhar and Maguire, this study finds that older patients with a history of healthcare utilization are at most risk for increased annual charges in the following year.[vi] Interestingly, Missing data elements are correlated to less costly patient populations. Astoundingly, the mere presence of chronic conditions does not significantly effect cost of patient care.

In a more holistic light, throughout the late 1990's, presidents, namely Nixon and Clinton, attempted to address the first pillar of cost control in healthcare by changing the payment mechanism in healthcare to managed,

capitated care rather than fee-for-service care. Changing the payment methodology would theoretically free physicians to manage patient care on an annual basis.[vii] However, the initial success of Health Maintenance Organizations (HMOs) on healthcare cost control was largely overshadowed by the sketchy money dealings in both the government and private sector alike which drove some healthcare providers to file bankruptcy.[viii] The failure was mainly a result of physician's inability to manage their patients' health without changing patient behavior.

The passing of the ACA, which has successfully expanded access to around 18 million more Americans, as well as the HITECH act, which is moving provider groups towards fee-for-value payment models starting in 2017, is spawning health system mergers across the industry. A crossroads is approaching in the next 5-6 years, where health systems are going to take on financial risk again like in the 90s. But this time the ACA is meant to promote the move towards the Accountable Care Organization (ACO) rather than the HMO. In the ACO model, providers will ideally have the data assets to sufficiently manage population health of their patient populations.[ix] Decision trees and random forest prediction models may be extremely useful in helping ACOs succeed where HMOs did not.

A 79% accuracy, although modest, instruments the potential for increased patient health through decision tree and random forest learning algorithms as applied to patient data. Future studies should delve into specific chronic condition types that may be more costly. Also, since past utilization was the chief indicator, additional data fields that granulize past utilization procedure types may further the accuracy and precision of this predictive model. This machine learning study represents one of the most complete, expansive examinations of population health with over 600,000 patients across 6 of the top 10 most populous states in the U.S.

---

[i] Singh, S. P., & Sawhney, T. G. (2006). Predictive analytics and the new world of retail healthcare. *Health management technology*, *27*(1), 46.

[ii] Ibid.

[iii] Maguire, J., & Dhar, V. (2013). Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: data-driven predictive analytics in healthcare. *Health Systems*, *2*(2), 73-92.

[iv] Bradley, P. (2012). Predictive analytics can support the ACO model.*Healthcare financial management: journal of the Healthcare Financial Management Association*, *66*(4), 102-106.

[v] Lin, J. H., & Haug, P. J. (2008). Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of biomedical informatics*, *41*(1), 1-14.

[vi] Maguire et al. (2013). 73-92.

[vii] Barr, Donald A. Introduction to U.S. Health Policy: the Organization, Financing, and delivery of health care in America. 2nd edition. Baltimore, MD: The Johns Hopkins University Press. Pages 121-122.

[viii] Ibid.

[ix] Gottlieb, Scott. (2015). How Many People Has Obamacare Really Insured? Forbes.