“Using Extreme Gradient Boosting in Claims Data to Predict Future Costs in a Health First Colorado Population”

Colorado ACCESS

coaccess.com

800-511-5010

customer.service@coaccess.com

11100 East Bethany Drive
Aurora, CO 80014

01 25-112 0121A

Suggested Citation: Davis, J. M., Ahle, J., Suleta, K. Using extreme gradient boosting in claims data to predict future costs in a Health First Colorado Population. January 2021. Colorado Access.

Executive Summary of Findings

- Predicting future costs in a Medicaid population could help with the management of high utilizing members and has been a goal in numerous projects.
- Here we develop a machine learning approach, extreme gradient boosting (XGBoost), using approximately 3,700 claims-based predictors as well as additional membership characteristics and demographics. We use utilization history over three quarters to predict costs incurred in a fifth quarter.
- The algorithm performs well with an R^2 value of 0.786 in a population of consistent utilizers. A robust Shapley additive explanation importance analysis suggests that previous costs are far more important to predicting future costs than other claims-based information in these data.

Background

Health care in the United States is costly. In 2018, health care costs reached $3.6 trillion and costs are projected to grow at an annual average rate of 5.5% per year between 2018 and 2027.[1,2] Health care costs may also represent nearly 20% of the gross domestic product by 2027.[2] Medicaid, the public insurance program funded at the federal and state levels in the United States, is the largest payer in the country and cost $597.4 billion in 2018.[3,4]

Claims-based analyses have been used extensively in the past to inform on and predict utilization costs. An early attempt to group claims into important cost categories for risk stratification was conducted in the development of the Diagnostic Cost Groups (DCG) in 1989.[5] Since the development of DCGs, numerous groups

have used claim categorizations in risk adjustment and cost forecasting. Recently, however, some reports have attempted a comprehensive machine learning (ML) solution in cost forecasting using claims and other data sources and have shown good performance.[6,7]

ML has been used to help identify and predict patterns in diverse health care data.[8] ML allows the synthesis of large amounts of predictors and complex non-linear patterns and interactions in predicting outcomes. However, while ML approaches are useful in this regard, they can also lead to puzzling interpretations regarding the relevance of predictors. Only recently has this improved, in part by the development of Shapley additive explanations (SHAP), which allow for a more robust understanding of the relevance of predictors.[9,10] Insights and efficiencies may be gained through a better understanding of claims-based predictors of future costs derived from ML. Here we present important claims-based predictors of costs through SHAP analyses applied to an algorithm developed with Extreme Gradient Boosting (XGBoost).[11]

Methods

This report combines Health First Colorado (Colorado's Medicaid Program) claims and member enrollment data from Colorado Access (COA), a nonprofit health plan focused on public insurance. COA is the Regional Accountable Entity (RAE) for two regions in Health First Colorado. These regions encompass Adams, Arapahoe, Denver, Douglas, and Elbert counties.

The timeframe of this project included claims data collected over five quarters between December 2018 through the end of February 2020. Data utilized for this project were created from claim aggregations over time, health utilization indicators based on claims, and demographic information. Claims data included quarterly counts of the first International Classification of Diseases (ICD) code, as well as Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) codes grouped using the Clinical Classifications Software (CCS) methodology from the Healthcare Cost and Utilization Project.[12] In addition, pharmacy claims grouped by highest level Generic Product Identifier (GPI), as well as revenue codes and place of service (POS) codes, were tallied for each individual quarterly. Revenue codes and place of service codes were not grouped due to fewer levels. Important cost predictors assembled included adjusted annual paid amounts (total amount paid in the last 12 months, annualized for members with inconsistent enrollment), previous month paid amounts, and previous quarters' paid amounts. Numerous paid amounts over time were included because historical paid amounts have been shown to be powerful predictors of future paid amounts.[13]

Data were assembled in a lagged manner wherein three quarters were used to predict the summed paid amounts per member in a fifth quarter. The fourth quarter of data were skipped to simulate delays commonly experienced in Health First Colorado claim submissions and processing. Data were assembled to stratify by consistent utilizers (i.e., individuals who had at least one claim per quarter over the timeframe). This stratification was undertaken to improve algorithm performance as seen in other reports focusing on high-cost utilizers[6,14] and to maximize limited computational resources. Further, stratification in consistent utilizers is important from the standpoint of the dynamic and transient population of Health First Colorado where large proportions of membership can change on an annual basis. Additionally, Medicaid management organizations often do not have extensive computational resources dedicated to ML, and a component of this project is to demonstrate reasonable application of ML in simple computational systems. The final data set included 146,413 individuals with 3,768 predictors. The project was conducted in R version 4.0.2 (https://cran.r-project.org/) on a Windows 10 desktop machine with 6 cores, 12 logical processors, and 128 GB of RAM.

Given the sparse nature of the data and the possibility of important but rare combinations, 139,093 observations (95%) of data were randomly selected for regression algorithm training by 5-fold cross-validation using the caret function in R.[15] The remaining 5% of data were used as a final testing data. Extreme gradient boosting (XGBoost) was trained given its exceptional computational performance as well as increased performance common to decision tree-based algorithms in structured data. A Shapley additive explanation (SHAP) variable importance evaluation was also conducted (Table 1).[16] SHAP importance has been shown to be robust to inconsistencies that arise in other measures of variable importance that can result from varying the order of features used in decision tree methodologies.[9,10] Finally, a follow-up XGBoost algorithm was trained with only costs and home- and community-based services.

Results

The population was mostly female (58.3%), mean aged 29.5 years, of mutually defined race/ethnicity (42.6%) and predominantly from Adams County (26.0%). Nine and a half percent did qualify for home- and community-based services.

Adjusted annual costs were the most important predictor followed by other cost variables. Eligibility for home- and community-based services was also notable; however, cost variables are substantially more relevant than all other features.

[Table 1 Mean SHAP values of the top 10 predictors]

Adjusted paid past year	1676
Quarter 3 total paid amount	917
Paid amount in the previous month	655
Home and community-based services	199
Quarter 2 total paid amount	180
Quarter 3 POS code 12: patient home	100
Quarter 1 total paid amount	62
Quarter 1 POS code 12: patient home	53
Number of risk indicators	51
Quarter 3 ICD10 group SYM016: other general signs and symptoms	42

[Table 2 Performance of Extreme Gradient Boosting algorithm]

	RMSE	R-squared
Extreme Gradient Boosting	3901	0.786
Extreme Gradient Boosting*	3954	0.782

*XGBoost that included only costs and home- and community-based services as predictors

XGBoost with only cost and home- and community-based services performed comparably to the primary XGBoost. Both algorithms performed well with R^2 values greater than 0.78.

Discussion

These analyses were able to capture extensive claims-based multivariate time series data with ICD codes, procedure codes, pharmacy codes, revenue codes, and place of service codes. From these data we applied XGBoost as well as a SHAP analysis. The SHAP analysis is robust to order of feature inclusion in tree methodologies and is an important advancement lending increased interpretability to what is frequently termed a black box of ML algorithms. Through these analyses, we have demonstrated that algorithm performance in consistent utilizers can be very good with minimal historical cost information and may not be improved with additional extensive claims-based information. This supports findings suggesting that claims-based information may only contribute to modeling of high-cost utilizers.[14]

While previous work has indicated that prior health care costs are a strong predictor of future costs,[17] this project suggests better performance than has been found in other literature. Though global comparisons among similar literature using R^2 alone should be taken with caution due to different populations and study objectives, the XGBoost R^2 of 0.79 suggests very good performance. One review found that most algorithms' R^2 values were clustered around 0.2, with the highest at 0.47.[18] The increased R^2 in this project is likely due to focusing solely on predicting costs of consistent utilizers of the health care system, or members who had at least one paid claim per quarter. Medicaid membership is traditionally transient, with members frequently losing and subsequently regaining

their enrollment status due to changes in income or other factors, making predictions difficult due to inconsistent utilization data.

Finally, we demonstrate that high-performing ML algorithms can be generated with minimal computational and data resources when stratification and a focus on costs are employed. To our knowledge SHAP analyses have not been conducted with claims data in these populations and they show the importance of costs in these data appear to far outweigh the importance of claims categorizations in near-term forecasting. Medicaid management organizations can use this information to their benefit when attempting to prevent future unnecessary costs and conduct better managed care for high-utilizing populations.

References

1. National health expenditures: aggregate and per capita amounts, annual percent change and percent distribution: selected calendar years 1960-2018. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical (2015).
2. National health expenditures projections 2018-2027. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsProjected (2019).
3. CMS Roadmaps for the Traditional Fee-for-Service (FFS) Program: Overview. file:///T:/Evaluation_and_Research/White%20Papers/HCU%20Predictive%20Model/Articles/RoadmapOverview_OEA_1-16.pdf.
4. NHE Fact Sheet: Historical NHE, 2018. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet (2019).
5. Ash, A., Porell, F., Gruenberg, L., Sawitz, E. & Beiser, A. Adjusting Medicare capitation payments using prior hospitalization data. Health Care Financ. Rev. 10, 17–29 (1989).
6. Ng, S. H. X. et al. Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore. BMJ Open 10, (2020).
7. Yang, C., Delcher, C., Shenkman, E. & Ranka, S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed. Eng. Online 17, (2018).
8. Wiens, J. & Shenoy, E. S. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin. Infect. Dis. 66, 149–153 (2018).
9. Lundberg, S. M. et al. Explainable AI for Trees: From Local Explanations to Global Understanding. ArXiv190504610 Cs Stat (2019).
10. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in Advances in Neural Information Processing Systems 30 (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
11. XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://dl.acm.org/doi/10.1145/2939672.2939785.
12. HCUP-US Home Page. https://www.hcup-us.ahrq.gov/.
13. Morid, M. A. et al. Healthcare cost prediction: Leveraging fine-grain temporal patterns. J. Biomed. Inform. 91, 103113 (2019).
14. Bertsimas, D. et al. Algorithmic Prediction of Health-Care Costs. Oper. Res. 56, 1382–1392 (2008).
15. Kuhn, M. et al. caret: Classification and Regression Training. (2020).
16. Liu, Y. liuyanguu/SHAPforxgboost. (2020).
17. Morid, M. A., Kawamoto, K., Ault, T., Dorius, J. & Abdelrahman, S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. AMIA. Annu. Symp. Proc. 2017, 1312–1321 (2018).
18. Kuo, R. N. et al. Predicting healthcare utilization using a pharmacy-based metric with the WHO's Anatomic Therapeutic Chemical algorithm. Med. Care 49, 1031–1039 (2011).