Predicting High-cost Health Insurance Members through Boosted Trees and Oversampling: An Application Using the HCCI Database

Brian Hartman[a,*], Rebecca Owen[b], Zoe Gibbs[a]

[a] *Department of Statistics, Brigham Young University, Provo, UT, USA*
[b] *HealthCare Analytical Solutions, Inc., Bend, OR, USA*

Abstract

Using the Health Care Cost Institute data (approximately 47M members over 7 years), we examine how to best predict which members will be high-cost next year. We find that cost history, age, and prescription drug coverage all predict high costs, with cost history being by far the most predictive. We also compare the predictive accuracy of logistic regression to extreme gradient boosting and find that the added flexibility of the extreme gradient boosting (xgboost) improves the predictive power. Finally, we show that with extremely unbalanced classes (because high-cost members are so rare) oversampling the minority class provides a better xgboost predictive model than undersampling the majority class or using the training data as is. Logistic regression performance seems unaffected by the method of sampling.

1. Introduction

A small proportion of members are responsible for a large majority of the total healthcare costs. While most people use very few services, mainly preventive care or minor acute care, and others are regular consumers, but at a moderate cost, nearly 75% of all healthcare expenditures are made by only 17% of users (McWilliams and Schwartz, 2017). This high-cost care can be attributed to four main groups.

1. Completely unexpected - burns or serious car accidents or the transition of a normally mild disease into a crisis due to an unexpected and unavoidable situation, such as the development of encephalitis from a case of West Nile Virus.
2. Lack of due care and caution, as well as some terrible luck, for example septicemia.
3. Expected but not necessarily predictable, like cancer care.
4. Chronic disease that has worsened in severity, so that she is fighting for her life after years of debility.

Due to the contribution to costs of this small segment of the population, there is considerable interest to understand what portion of it can be predicted. The portion that is entirely random, and rare, may be estimated by a distribution based on large population studies. It is the portion that could be estimated using a predictive model based on the characteristics of the population that is of great interest because it would allow for some predictions in future costs of a specific population, as well as identifying people for interventions and additional care. While risk adjustment models are good at predicting average costs of care for a category of people, they are still not effective at identifying particular people who may be at risk for very high claim costs in the near future (Hileman et al., 2016; Hileman and Steele, 2016).

This research explores the types of models that will identify individuals who are most likely to exceed a high cost threshold based on a number of characteristics available in the most common information source: administrative claims data. While additional information from chart review and clinical recommendations would be helpful to identify members at risk, this data is often difficult to incorporate into the actuarial

*Corresponding author
Email addresses: hartman@stat.byu.edu (Brian Hartman), rebecca@hcasolutions.com (Rebecca Owen), zgibbs8@gmail.com (Zoe Gibbs)

studies for trend and pricing work. This work seeks to add another tool to the risk quantification process for members whose costs form a large part of the overall costs of care as well as a significant contributor to the force of trend.

Many authors have examined the issue of high-cost claimants from different directions. A first group explored the common characteristics of high-cost members. Zook and Moore (1980) looked at 2238 patients (of which 13% were high-cost) and found that smoking and drinking were much more prevalent in the high-cost group than the low-cost one. Schroeder et al. (1979) found that very few (17%) of the high-cost claimants suffered from an actual medical catastrophe; most had chronic conditions. Joynt et al. (2013) found that only a small portion of the total spending for high-cost members was due to preventable acute care. Zulman et al. (2015) showed that multimorbidity is common among high-cost members of the U.S. Veterans Affairs Health Care System. They suggest that interventions are needed to help those members better manage multiple conditions. In order to effectively assign these interventions, we need to predict who will likely be high-cost.

Another group looked at which covariates are most likely to predict high-cost members. Garfinkel et al. (1988) used the National Medical Care Utilization and Expenditure Survey to look at predictors of high-cost patients. They found that health status, followed by economic factors best predict high-cost members. Meenan et al. (2003) compared many risk-adjustment models available at the time to determine which are the best at predicting high-cost patients. Fleishman and Cohen (2010) analyzed the Medical Expenditure Panel Survey (MEPS) and compared a risk score (diagnostic cost group) with a count of chronic conditions on their ability to predict which members would be in the highest cost decile the following year. They also checked whether self-rated health status and functional limitations improved predictions. They found that the risk score was the best predictor. After controlling for the risk score, the number of chronic conditions, self-reported health status, and functional limitations were significantly associated with future high-costs.

A final group, whose work most closely aligns with our paper, focus on developing optimal methodology for predicting high-cost members given available covariates. There are two main ways to predict who will be high-cost. Predicting the actual costs for the member will give an entire predictive distribution. Then calculating the probability of any cost, or of exceeding any threshold is trivial. Bayesian hierarchical models (Fellingham et al., 2005), Bayesian nonparametric regression (Fellingham et al., 2015; Hong and Martin, 2017; Richardson and Hartman, 2018), two-part models (Rosenberg and Farrell, 2008; Frees et al., 2011, 2013), and machine learning models (Duncan et al., 2016; Robinson, 2008; Moturu et al., 2007, 2009) can be used to solve this problem. Accurate prediction of probabilities in the (notably heavy) tails can be difficult due to the lack of extreme data and will be largely dependent upon model assumptions. For those reasons, we focus on predicting the probabilities that members will exceed certain thresholds. While the predictions are not as detailed as those obtained from the total cost models mentioned earlier, they are not as dependent on model assumptions and focus on a simpler question which the sparse extreme data are better able to answer. Exceedance probabilities also naturally answer the question of how likely certain members are to benefit from intervention, both in cost and member outcome.

2. Data

Our data was gathered by the Health Care Cost Institute. It consists of member information from three of the largest health insurers in the United States. When we performed our analysis, they had data for each year 2009-2015. The number of members in each year are listed in Table 1.

The variables we are interested in for our analysis are described in Table 2. We divide the members into five groups based on their allowed, adjudicated costs for the year (<100K, 100K-250K, 250K-500K, 500K-1M, >1M). The vast majority of the members had less than $100,000 in total claims each year. To understand the rarity of the group we are exploring, Table 3 shows the number of members in each high-cost group. As is readily apparent, there are not many members in the extremely high-cost group. This is another reason to focus on the probability of exceeding a certain threshold, rather than attempt to estimate a predictive distribution of costs for each member. We are essentially taking the role of an intervention manager and trying to find those members which are most likely to be high-cost. The proportion of all members in each of the high-cost groups have also increased every year.

Year	Number of Members
2009	48,511,544
2010	47,539,751
2011	46,193,435
2012	46,544,359
2013	47,351,996
2014	48,087,209
2015	47,782,320

Table 1: Number of members in each dataset

Variable Name	Description
Z_PATID	Member ID number
RX_CVG_IND	Prescription drug coverage indicator (1 if the member has coverage). If 1, the pharmacy costs for the year are included in the total allowed costs below.
GDR	Gender (1 for male, 2 for female)
AGE	Age in years
MKT_SGMNT_CD	Market segment code (I-Individual market, G-Individual group conversion, L-Large, S-Small, O-Other) For inference, we focus only on the individual market (INDV_FLAG), but in prediction we use all segments.
CAT	Total allowed, adjudicated cost for the year, divided into five groups (<100K, 100K-250K, 250K-500K, 500K-1M, >1M)
CATLESS1	Total allowed, adjudicated cost for the member one year ago, divided into five groups (<100K, 100K-250K, 250K-500K, 500K-1M, >1M)
CATLESS2	Total allowed, adjudicated cost for the member two years ago, divided into five groups (<100K, 100K-250K, 250K-500K, 500K-1M, >1M)

Table 2: Variable names and descriptions

When predicting whether a member will be high-cost in a certain year, we only use data available at that time (similar to how the analysis will be done in practice). We will use data from the previous two years to predict if the member will be high-cost in the following year. For example, to predict whether the member will be high-cost in 2012, we will use data from 2010 and 2011. Because we have data from 2009-2015, we predict each member in 2011-2015. For each prediction year, we only use those members for which we have at least some data for the year in question and the previous two. That reduces the sizes of our analysis datasets to those shown in Table 4.

Reducing our dataset to only those who were members for at least part of each of three sequential years impacts our data (and therefore our inference) in several ways. First, there are no members under the age of 2 in our prediction datasets. Partially because of those missing infants, the median age of the members in our analysis dataset is about nine years older than that of those not in our set (39 vs. 30). Further, much of the lifetime medical spending occurs in the final year of life, so those who were expensive and then passed away in previous years will not be included in our analysis dataset. About 25% of the high cost members (>100K) from any year are not in the dataset in the following year. Of those in our analysis dataset, around 50% have prescription drug coverage, whereas of those not in our analysis dataset about 60% have coverage. Additionally, there are about twice as many members in the individual market among those not in our analysis dataset (about 8% to about 4%), potentially due to people moving to the Affordable Care Act's exchanges and our data being unable to connect that person in the two different providers. Most importantly, the proportion of high-cost members is about the same between the two groups, except for a few more people (one or two per 100K members) above 1M in the group not in our analysis dataset.

Year	100K-250K	250K-500K	500K-1M	>1M
2009	96,554	17,738	4,162	661
2010	100,812	18,162	4,393	706
2011	108,965	20,375	4,773	841
2012	117,325	22,393	5,250	941
2013	126,099	24,275	5,458	998
2014	135,050	26,018	5,749	1,030
2015	147,220	28,425	6,517	1,200

Table 3: Number of members in each high-cost group

Prediction Year	Sample Size
2011	25,954,734
2012	26,539,732
2013	27,061,494
2014	26,425,810
2015	25,199,632

Table 4: Sample size of each three-year dataset

3. Methods

For inference, we fit separate logistic regression models to four different thresholds of high cost, greater than 100K, 250K, 500K, and 1M in claims. We are interested in the parameter estimates and whether they change over time. Looking at the parameter estimates over time will allow us to see both how consistent the estimates are and to notice any temporal changes or patterns (possibly due to the Affordable Care Act).

For prediction, we will compare two separate models, logistic regression and extreme gradient-boosted classification trees (Chen and Guestrin, 2016), also known as xgboost. Classification trees attempt to model which observations are likely to be high-cost by repeatedly splitting the dataset on different explanatory variables, trying to make the resulting subsets of observations as similar as possible (containing mainly positive or negative cases) while preventing overfitting. Xgboost refines standard classification trees by fitting new trees to the residuals resulting from earlier trees. The effectiveness of an xgboost model largely depends on the hyperparameter settings which we discuss and optimize later in this section. For our particular application, we fit all of the models in R within the HCCI enclave. With the dataset being so large, the models take between 5-30 minutes to fit. That will limit the amount of cross-validation we can do as we are optimizing the hyperparameters.

When we compare the predictive accuracy of the two models, we fit the model to one year (say using 2010 and 2011 to predict 2012), and use that model to fit the following year (2013, using 2011 and 2012). This will show which model is superior in a realistic situation. This is better than dividing each year into a training and test set and comparing model accuracy that way.

Classification can be difficult when the positive class is extremely rare, as it is in our case. To help mitigate that issue, we will train the 2012 models on three different datasets.

- Standard: The original 2011 data (say 1,000 high-cost members and 1,000,000 low-cost members for illustration)
- Under: A dataset with the 2011 low-cost members undersampled, making an equal number of high- and low-cost members. We randomly select (without replacement) 1,000 of the 1,000,000 low-cost members to be in the training set. This means that we have a training sample of 2,000 members.

- Over: A dataset with the high-cost members oversampled, again making an equal number of high- and low-cost members. We randomly select (with replacement) a sample of 1,000,000 from the 1,000 high-cost members. In this case, our training sample will include 2,000,000 members.

To tune the xgboost models, we adjust the following five hyperparemters:

- Maximum tree depth, ranging between (3, 10) - maximum number of branch levels in any tree. A higher number here makes it more likely that an individual tree is overfit.
- Minimum child weight (1, 10) - This parameter tells the tree-building process when to stop. If splitting a node would make a child have less weight than this parameter, then the process stops. The larger this value, the simpler the trees will be.
- Subsample (0.5, 1) - Proportion of the total training set used to build each tree. A smaller value will help to prevent overfitting.
- Column Sample by Tree (0.5, 1) - Proportion of all the possible covariates used to build each tree. A smaller value helps to prevent overfitting.
- Eta (0,1) - The learning rate. A higher eta will speed up convergence, while a lower eta may make the convergence more precise.

We created four different xgboost models. The first (untrained) uses default values for each of the above hyperparameters. The next model (trained1) starts with the hyperparameters from the untrained model and then compares it to ten different possible settings, randomly drawn from the set of possible hyperparameters (in parentheses in the list above). The settings are compared through 3-fold cross-validation and by choosing the set of hyperparameters which maximizes the AUC in the cross-validation. This is done only with the data available through 2011, making sure that the optimization does not include any of the data we are trying to predict. The following model (trained2) starts with the chosen hyperparameters in trained1, and then compares that to ten additional randomly drawn possible sets. The final model (trained3), follows the same pattern. The chosen hyperparameters are in Table 5.

Parameter	Untrained	Trained1	Trained2	Trained3
Maximum Tree Depth	6	3	5	5
Minimum Child Weight	1	9.77	2.98	9.26
Subsample	1	0.66	0.79	0.97
Column Sample by Tree	1	0.76	0.60	0.69
Eta	0.3	0.54	0.52	0.63

Table 5: Hyperparameter settings for the four models

Trained1 and Trained3 have similar hyperparameter settings, though the subsample rates are much higher for trained3. That could make overfitting more likely in trained3 than in trained1. Trained2 has a much smaller minimum child weight, which can also lead to more complicated trees and potential overfitting.

4. Results

We divided our results section into two parts. First, we examine the inference results gathered from the logistic regressions on each year 2011-2015, using the two previous years to help predict current year cost. Then, we will discuss the prediction results of the various models.

4.1. Inference

Because we are comparing four different definitions of high-cost (100K, 250K, 500K, and 1M, collectively referred to as thresholds) over five different prediction years (2011-2015) we will base the results on a baseline member who is a 35 year-old male with group coverage and no history of costs above 100K in the last two years.

4.1.1. Baseline Probabilities

In 2011, our baseline male has about a 0.0018 probability of having allowed costs of more than 100K (about 1 in every 550 members). The probabilities decrease as the threshold increases with the costs being 250 times more likely to be over 100K than over 1M (see Table 6).

Threshold	Probability	Relative to 100K
100K	0.001774	100%
250K	0.000345	19.4%
500K	0.000064	3.6%
1M	0.000007	0.4 %

Table 6: Predicted probability of a 35 year-old male being high-cost in 2011

These probabilities have also increased over time. Figure 1 shows the change in probability for our baseline male over time, relative to the probability in 2011. Notice that the probability being over 1M is growing most rapidly, followed by 500K, with 100K and 250K similar.

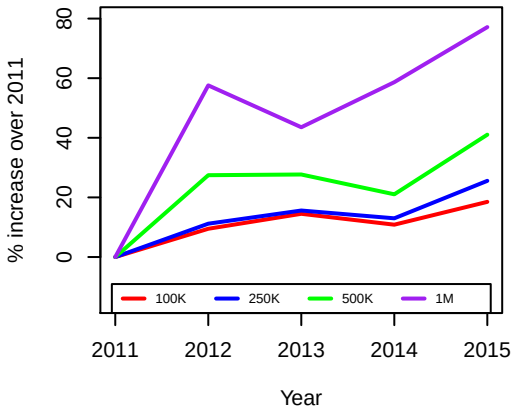

Figure 1: Increase in baseline predicted probabilities over time

4.1.2. Covariates

The rest of the coefficients (gender, age, prescription drug coverage, individual or group market, and high-cost history) have a relatively similar relationship with costs across analysis years, so we will combine the results from the five years into a single estimate and measure of uncertainty. Additionally, because the individual probabilities are so small, we express them relative to a baseline 35 year-old male with group coverage and no history of high claims. The relativities (and their confidence intervals) are displayed in Table 7.

Claims history has by far the largest effect on high-cost probability. A 35 year-old male member with over a million dollars in claims last year has almost a 40% chance of claiming more than 100K this year, up

	100K	250K	500K	1M
Age 35 Male (Baseline)	1.0 (0.8, 1.2)	1.0 (0.7, 1.3)	1.0 (0.6, 1.7)	1.0 (0.4, 2.8)
RX Coverage	1.2 (1.1, 1.5)	1.1 (0.8, 1.5)	1.0 (0.6, 1.7)	1.0 (0.3, 2.7)
Female	0.9 (0.8, 1.1)	0.8 (0.6, 1.1)	0.8 (0.5, 1.3)	0.8 (0.3, 2.2)
Individual Market	0.7 (0.6, 0.9)	0.8 (0.6, 1.0)	0.8 (0.5, 1.4)	0.8 (0.2, 2.5)
Age 15	0.5 (0.4, 0.5)	0.6 (0.5, 0.8)	0.9 (0.6, 1.5)	1.4 (0.6, 3.7)
Age 55	2.5 (2.1, 3.0)	2.0 (1.5, 2.7)	1.6 (0.9, 2.8)	1.3 (0.4, 4.2)
Age 75	2.6 (2.2, 3.1)	2.0 (1.4, 2.7)	1.5 (0.8, 2.8)	1.1 (0.3, 3.9)
100-250K last year	44.6 (38.2, 51.9)	61.0 (45.5, 81.6)	70.1 (41.5, 118.4)	70.0 (23.9, 205.2)
250-500K last year	93.4 (81.0, 107.2)	242.4 (184.0, 316.9)	313.7 (186.5, 524.0)	301.6 (102.0, 887.8)
500K-1M last year	139.2 (121.0, 159.0)	424.1 (325.5, 545.3)	1144.7 (696.9, 1836.6)	1414.0 (482.1, 4071.8)
1M+ last year	194.6 (162.0, 229.4)	626.6 (468.4, 817.3)	1893.0 (1137.7, 3019.5)	5458.6 (1864.6, 14898.3)
100-250K two years ago	6.8 (5.8, 8.1)	3.1 (2.3, 4.2)	2.1 (1.2, 3.5)	1.9 (0.7, 5.8)
250-500K two years ago	5.9 (5.0, 7.1)	4.6 (3.4, 6.3)	2.9 (1.7, 5.0)	2.1 (0.7, 6.3)
500K-1M two years ago	5.0 (4.0, 6.1)	4.8 (3.4, 6.6)	4.6 (2.7, 8.1)	3.4 (1.1, 10.4)
1M+ two years ago	6.7 (4.7, 9.5)	5.1 (3.2, 7.9)	4.9 (2.5, 9.5)	7.1 (2.1, 23.8)

Table 7: Predicted probability of being high-cost, relative to a 35 year-old male

from 0.2% if they had less than 100K in claims last year. Even having high claims two years ago has a big impact on the high-cost probability for this year. The rest of the covariates have small to negligible impacts.

4.2. Prediction

We compare the predictions from a logistic regression to those from gradient boosted trees, both with default hyperparameter settings and settings optimized through cross-validation. We compare the accuracy using the area under the ROC curve (AUC). The ROC curve plots the true positive rate against the false positive rate as the threshold changes. Therefore, the larger the AUC, the better the model discerns between high- and low-cost members. The AUC values are relatively constant across years, so we combined them into single estimates. Plots of the average AUC values across years are available in Figure 2.

For thresholds of 100K and 250K, all of the xgboost models significantly outperform logistic regression. Also, the sampling method doesn't seem to have much of an impact. As the number of positive cases decreases (for thresholds of 500K and 1M), oversampling outperforms the other two sampling methods for the xgboost models. This is less true for trained1, where undersampling works essentially as well as oversampling. The hyperparameter settings for Trained1 constrained the model to be relatively simple, possibly muting the benefit of the oversampling. For the other three xgboost models, undersampling is by far the worst method. Logistic regression performs equally well, regardless of sampling methodology, but in all cases is outperformed by each of the oversampled xgboost models.

5. Conclusion

In this paper, we use the Health Care Cost Institute data (approximately 47M members over 7 years) to examine how to best predict and describe high-cost members. Using a logistic regression, we find that cost history, age, and prescription drug coverage all predict high-costs, with cost history being by far the most predictive. In addition to the logistic regression model, we compare the predictive accuracy of extreme gradient boosting and find that the added flexibility of the extreme gradient boosting greatly improves the predictive power. Finally, we show that with our extremely unbalanced classes, oversampling the minority class provides better predictions than undersampling the majority class or using the training data as is.

There are many potential avenues of future work to explore. With the HCCI data, it would be very interesting to further explore the many relationships among the members (spatially, temporally, and hierarchically). We could look at the difference between those members with RX coverage and those without. It would also be interesting to try and quantify the impact of wellness programs, or attempt to explore the impact of the Affordable Care Act.

6. Acknowledgments

The authors acknowledge the assistance of the Health Care Cost Institute (HCCI) and its data contributors, Aetna, Humana, and UnitedHealthcare, in providing the claims data analyzed in this study. They are also grateful for the support of the Society of Actuaries which funded this work. Finally, they are grateful to Brad Barney and the anonymous referees for their suggestions and advice.

References

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 785–794. URL: http://doi.acm.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

Duncan, I., Loginov, M., Ludkovski, M., 2016. Testing alternative regression frameworks for predictive modeling of health care costs. North American Actuarial Journal 20, 65–87.

Fellingham, G.W., Dennis Tolley, H., Herzog, T.N., 2005. Comparing credibility estimates of health insurance claims costs. North American Actuarial Journal 9, 1–12.

Fellingham, G.W., Kottas, A., Hartman, B.M., 2015. Bayesian nonparametric predictive modeling of group health claims. Insurance: Mathematics and Economics 60, 1–10.

Fleishman, J.A., Cohen, J.W., 2010. Using information on clinical conditions to predict high-cost patients. Health services research 45, 532–552.

Frees, E.W., Gao, J., Rosenberg, M.A., 2011. Predicting the frequency and amount of health care expenditures. North American Actuarial Journal 15, 377–392.

Frees, E.W., Jin, X., Lin, X., 2013. Actuarial applications of multivariate two-part regression models. Annals of Actuarial Science 7, 258–287.

Garfinkel, S.A., Riley, G.F., Iannacchione, V.G., 1988. High-cost users of medical care. Health care financing review 9, 41.

Hileman, G., Mehmud, S., Rosenberg, M., 2016. Risk scoring in health insurance: A primer. Society of Actuaries .

Hileman, G., Steele, S., 2016. Accuracy of claims-based risk scoring models. Society of Actuaries .

Hong, L., Martin, R., 2017. A flexible bayesian nonparametric model for predicting future insurance claims. North American Actuarial Journal 21, 228–241.

Joynt, K.E., Gawande, A.A., Orav, E.J., Jha, A.K., 2013. Contribution of preventable acute care spending to total spending for high-cost medicare patients. Jama 309, 2572–2578.

McWilliams, J.M., Schwartz, A.L., 2017. Focusing on high-cost patients: the key to addressing high costs? The New England journal of medicine 376, 807.

Meenan, R.T., Goodman, M.J., Fishman, P.A., Hornbrook, M.C., OâĂŹKeeffe-Rosetti, M.C., Bachman, D.J., 2003. Using risk-adjustment models to identify high-cost risks. Medical care 41, 1301–1312.

Moturu, S.T., Johnson, W.G., Liu, H., 2007. Predicting future high-cost patients: a real-world risk modeling application, in: Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on, IEEE. pp. 202–208.

Moturu, S.T., Johnson, W.G., Liu, H., 2009. Predictive risk modelling for forecasting high-cost patients: a real-world application using medicaid data. International Journal of Biomedical Engineering and Technology 3, 114–132.

Richardson, R., Hartman, B., 2018. Bayesian nonparametric regression models for modeling and predicting healthcare claims. Insurance: Mathematics and Economics 83, 1–8.

Robinson, J.W., 2008. Regression tree boosting to adjust health care cost predictions for diagnostic mix. Health services research 43, 755–772.

Rosenberg, M.A., Farrell, P.M., 2008. Predictive modeling of costs for a chronic disease with acute high-cost episodes. North American Actuarial Journal 12, 1–19.

Schroeder, S.A., Showstack, J.A., Roberts, H.E., 1979. Frequency and clinical description of high-cost patients in 17 acute-care hospitals. New England Journal of Medicine 300, 1306–1309.

Zook, C.J., Moore, F.D., 1980. High-cost users of medical care. New England Journal of Medicine 302, 996–1002.

Zulman, D.M., Chee, C.P., Wagner, T.H., Yoon, J., Cohen, D.M., Holmes, T.H., Ritchie, C., Asch, S.M., 2015. Multimorbidity and healthcare utilisation among high-cost patients in the us veterans affairs health care system. BMJ open 5, e007771.

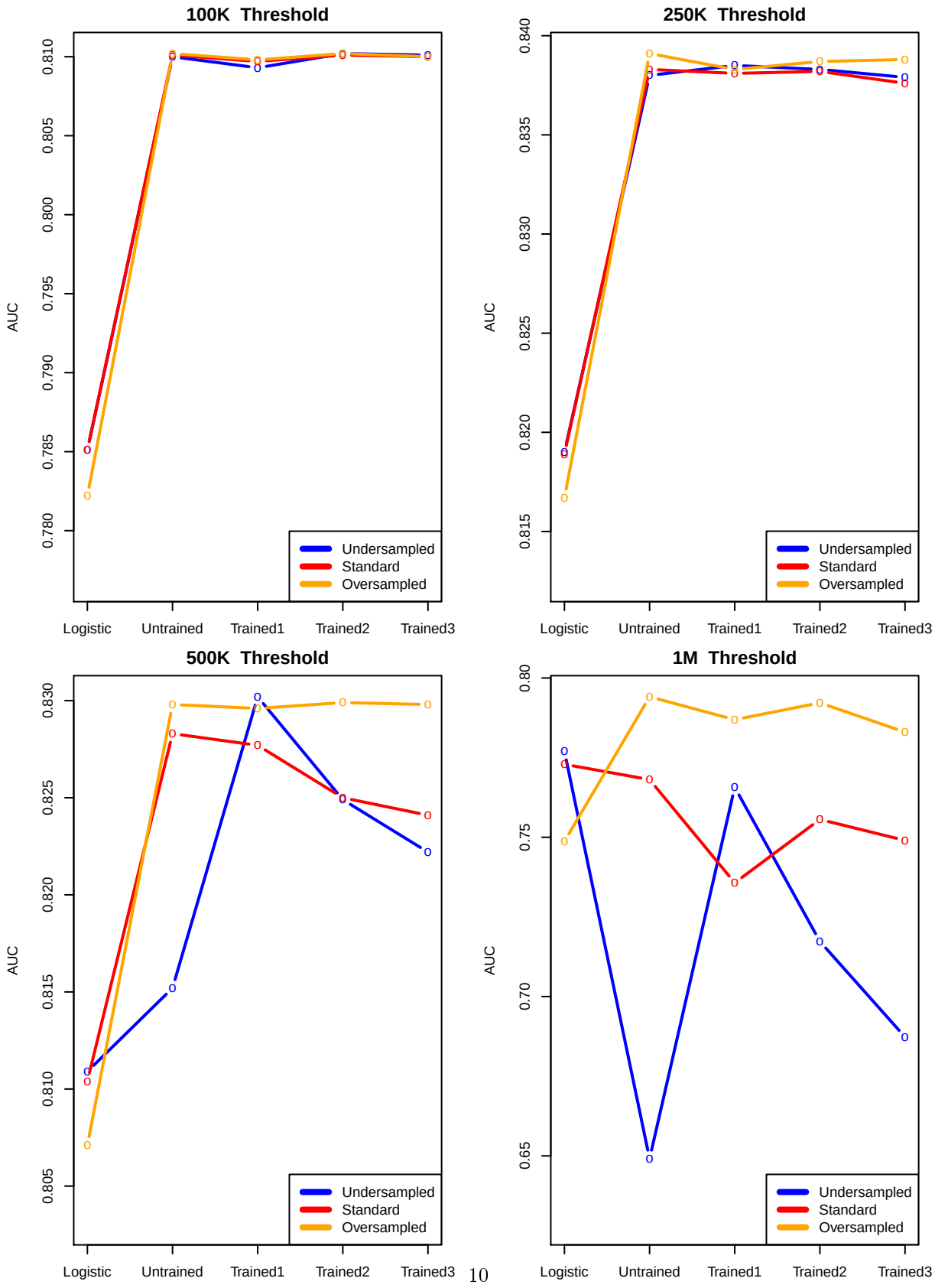

Figure 2: Average area under the ROC curve across years