Predicting High-cost Members in the HCCI Database

BRIAN HARTMAN, BRIGHAM YOUNG UNIVERSITY

JOINT WORK WITH REBECCA OWEN AND ZOE GIBBS

Acknowledgments

The Health Care Cost Institute (HCCI) and its data contributors, Aetna, Humana, and UnitedHealthcare, for providing the claims data analyzed in this study.

The SOA (led by Dale Hall) for funding this work.

Brad Barney for insightful comments and suggestions

Why is this important?

PREDICTING HIGH-COST MEMBERS IN THE HCCI DATABASE

Importance

Insurers and policymakers are very interested in predicting which members will be high-cost next year for:

- Assigning interventions (nurse, etc)
- High-risk pools
- General solvency
- Group rate renewals

Data

PREDICTING HIGH-COST MEMBERS IN THE HCCI DATABASE

Size of the HCCI Datasets

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

Explanatory Variables

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.
FEMALE	Gender (0 for male, 1 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)
CAT	Total allowed, adjudicated cost for the year, divided into five groups (<100K, 100K-250K, 250K-500K, 500K-1M, >1M)
CATLESS_1	CAT from one year prior
CATLESS_2	CAT from two years prior

Number of High-cost Members

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

Prediction Datasets

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

Inference

PREDICTING HIGH-COST MEMBERS IN THE HCCI DATABASE

Inference

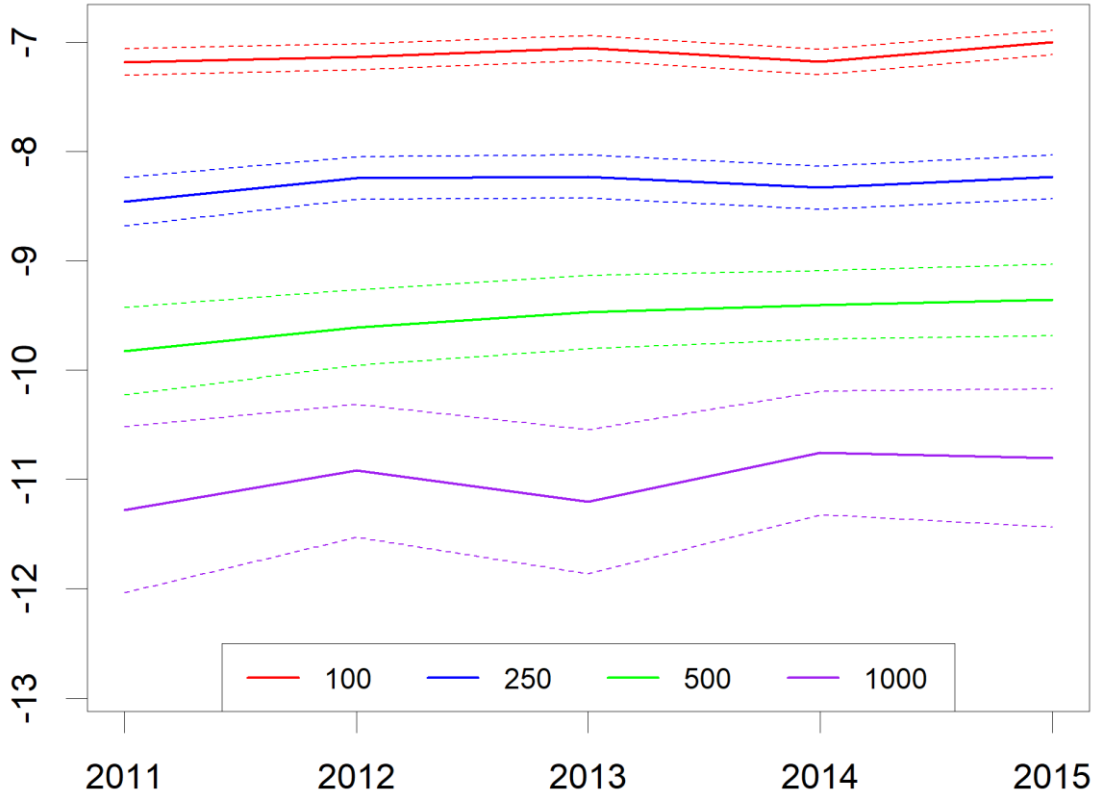
To help us understand what variables are really driving high-cost members, we fit logistic regressions to each

- Year
- High cost cutoff (cut)

We then compared the coefficient estimates (and confidence intervals) to look for trends.

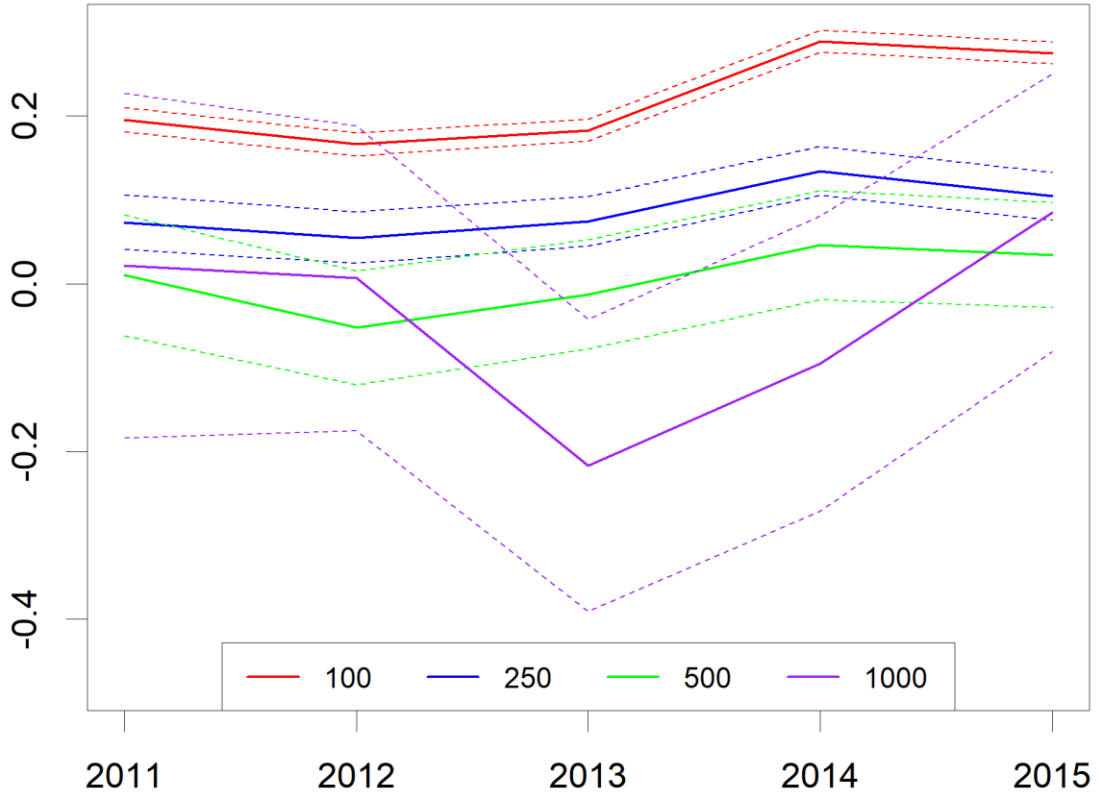
INTERCEPT

- All effects relatively constant between years
- In the correct order (a priori more likely to be above 100K than above 250K)



RX_CVG_IND

- Positive effect for 100
- Smaller positive effect for 250
- Not much of an effect for 500 or 1000

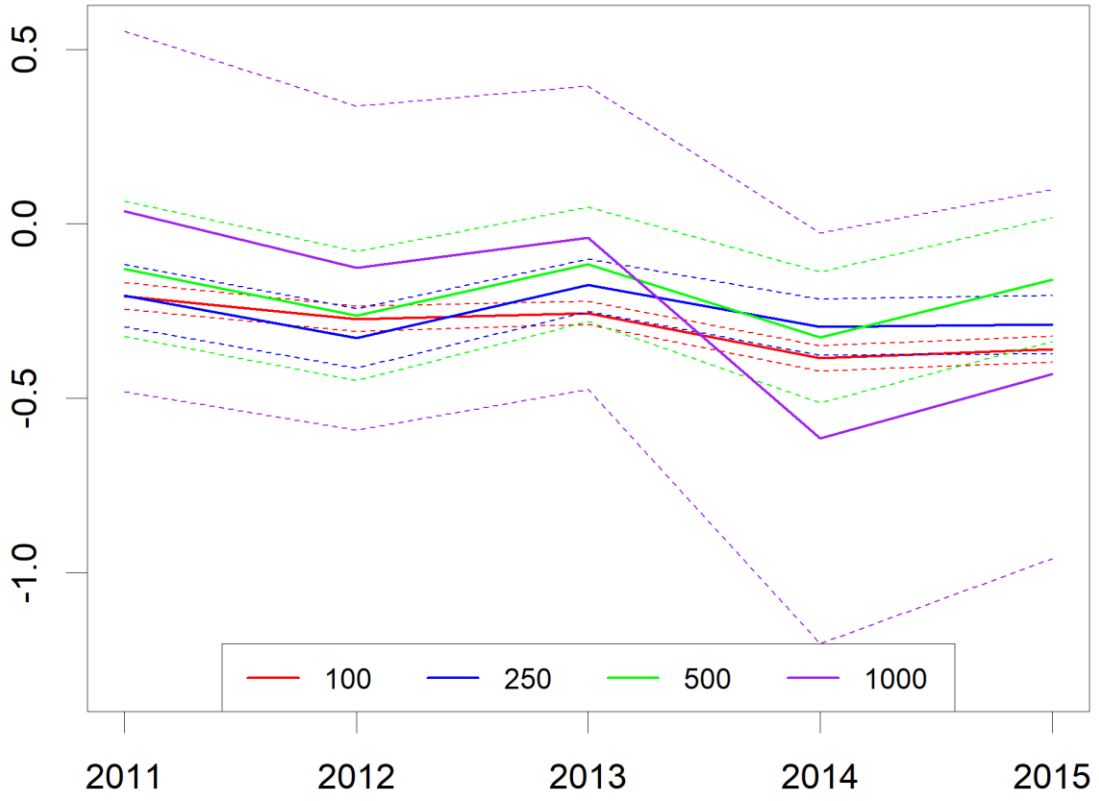


FEMALE

- Slight negative effect for 100
- Larger negative effect for 250, 500, and 1000

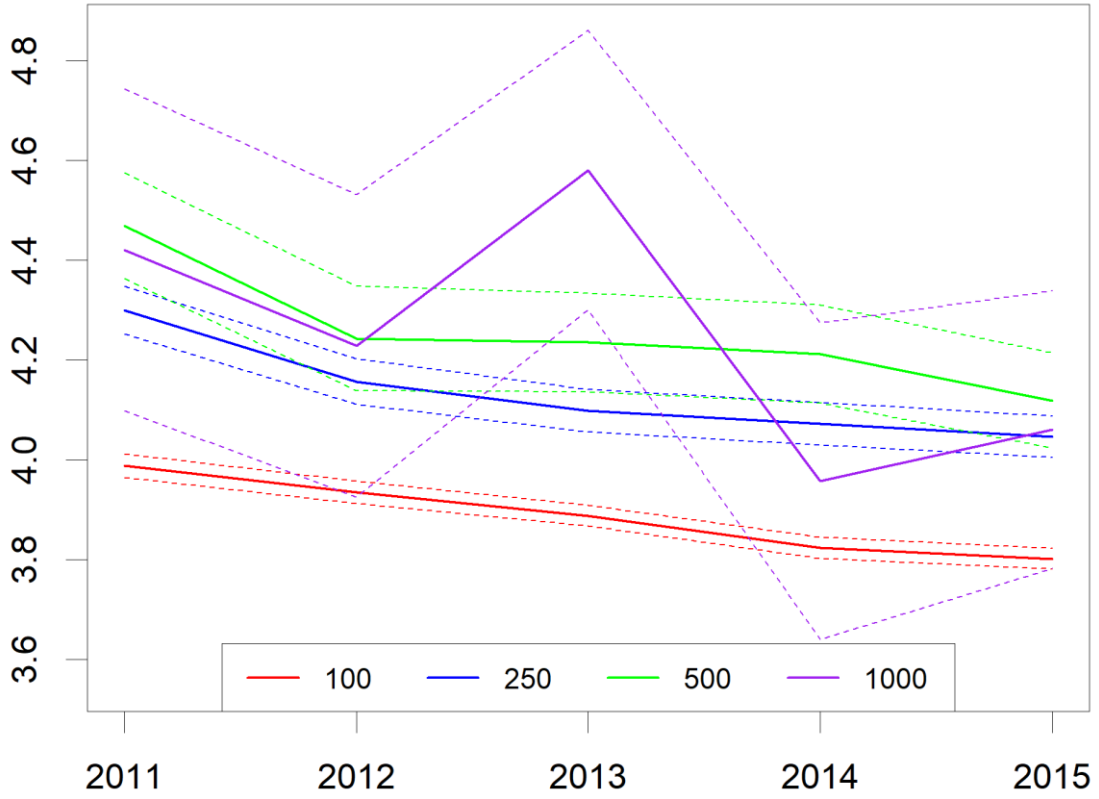
INDV_FLAG

- Slight negative effect 100 or 250
- No significant effect for 500 or 1000



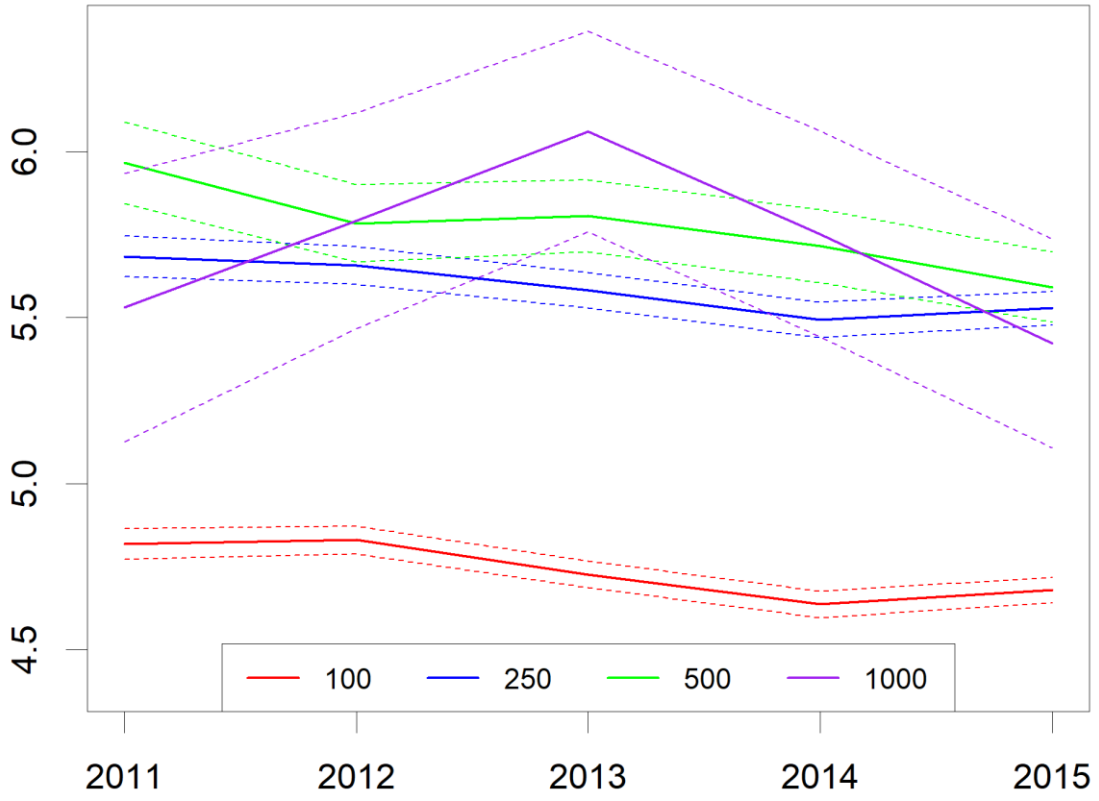
CATLESS1_100

- Large positive effect for all cuts.
- 100, 250, and 500 are in order from smallest to largest effect.
- Much larger uncertainty in 1000.



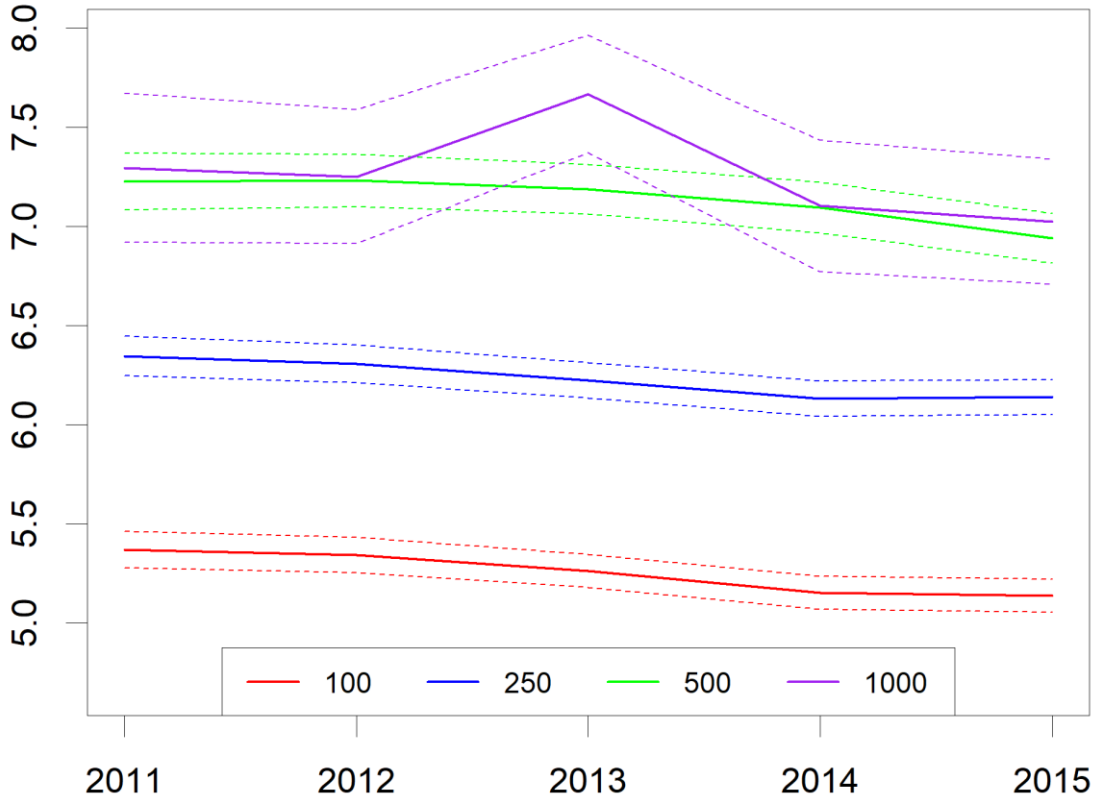
CATLESS1_250

- Larger effects than CATLESS1_100
- 100, 250, and 500 are in order from smallest to largest effect.
- Much larger uncertainty in 1000, though definitely a larger effect than for 100.



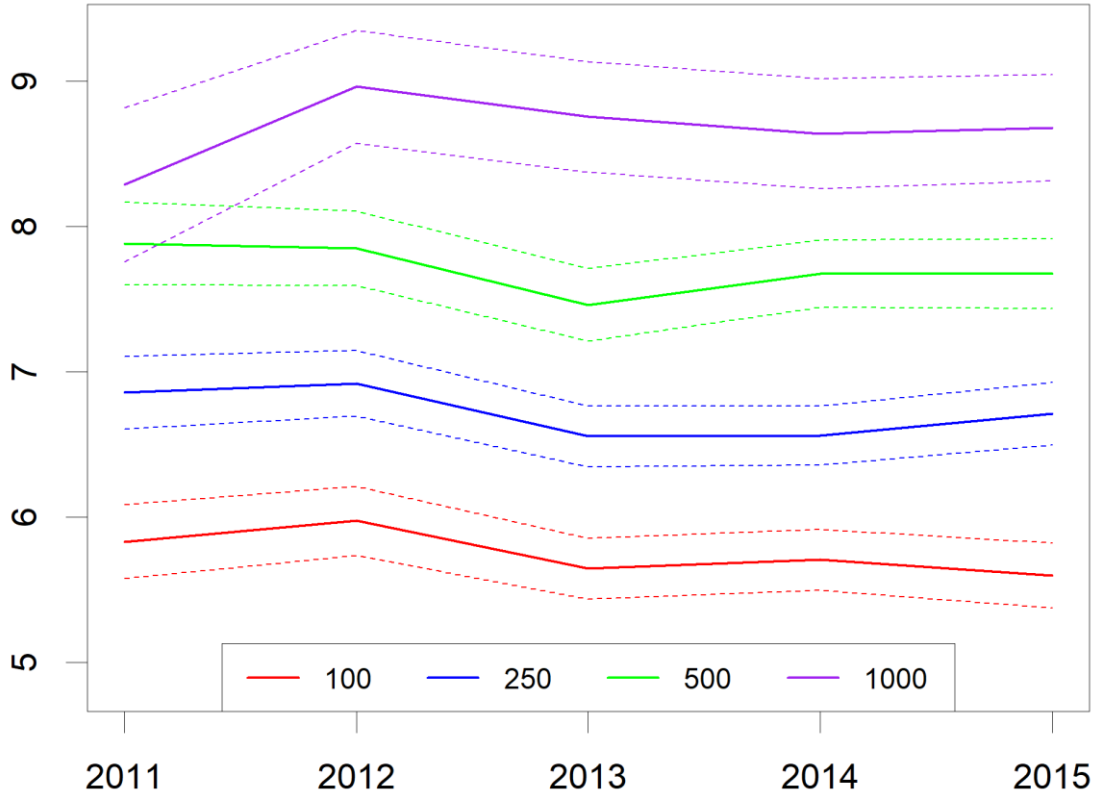
CATLESS1_500

- Continued increased separation.
- Stronger effects across years and cuts.



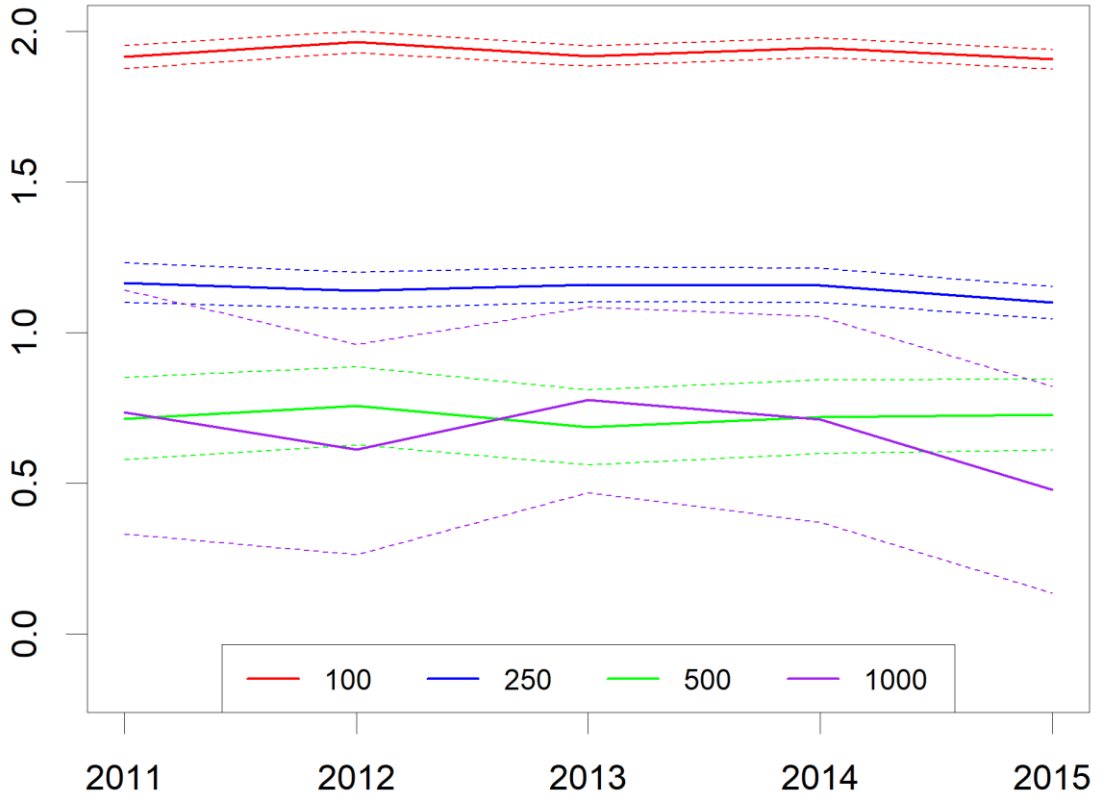
CATLESS1_1000

- Largest separation
- Largest effects
- Increased standard error



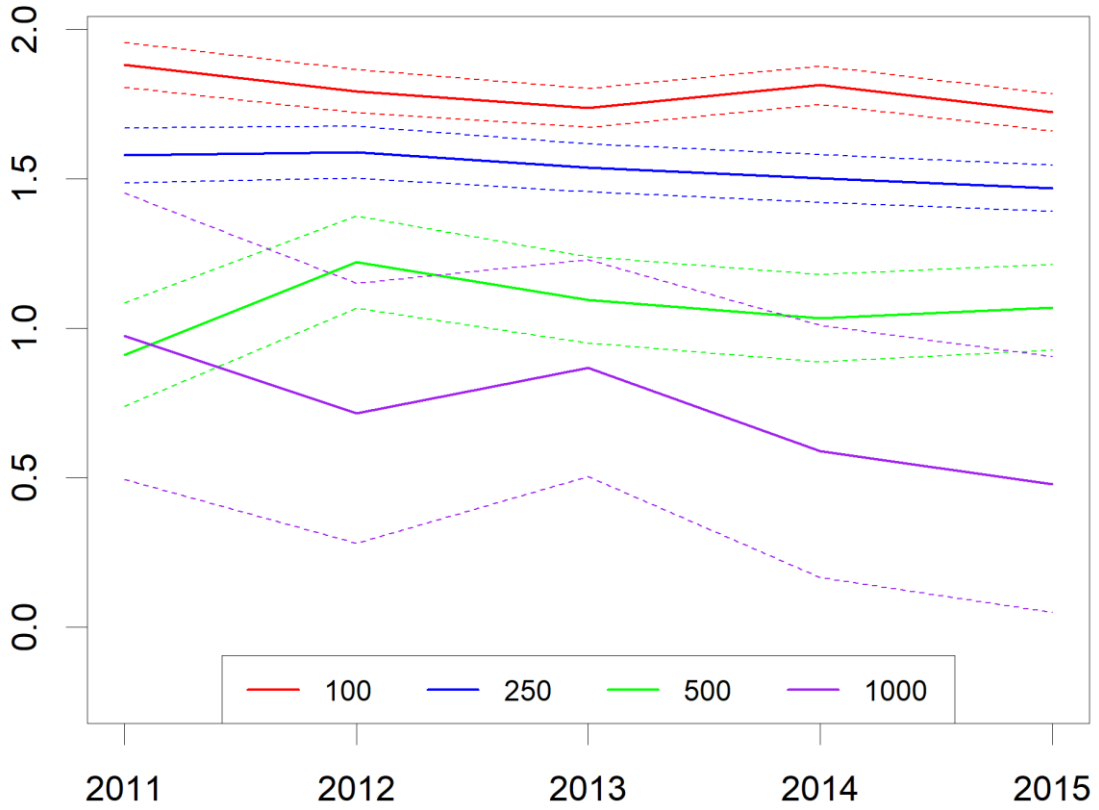
CATLESS2_100

- Biggest impact on 100
- Next largest impact on 250
- All impacts significantly smaller than those for CATLESS1_100



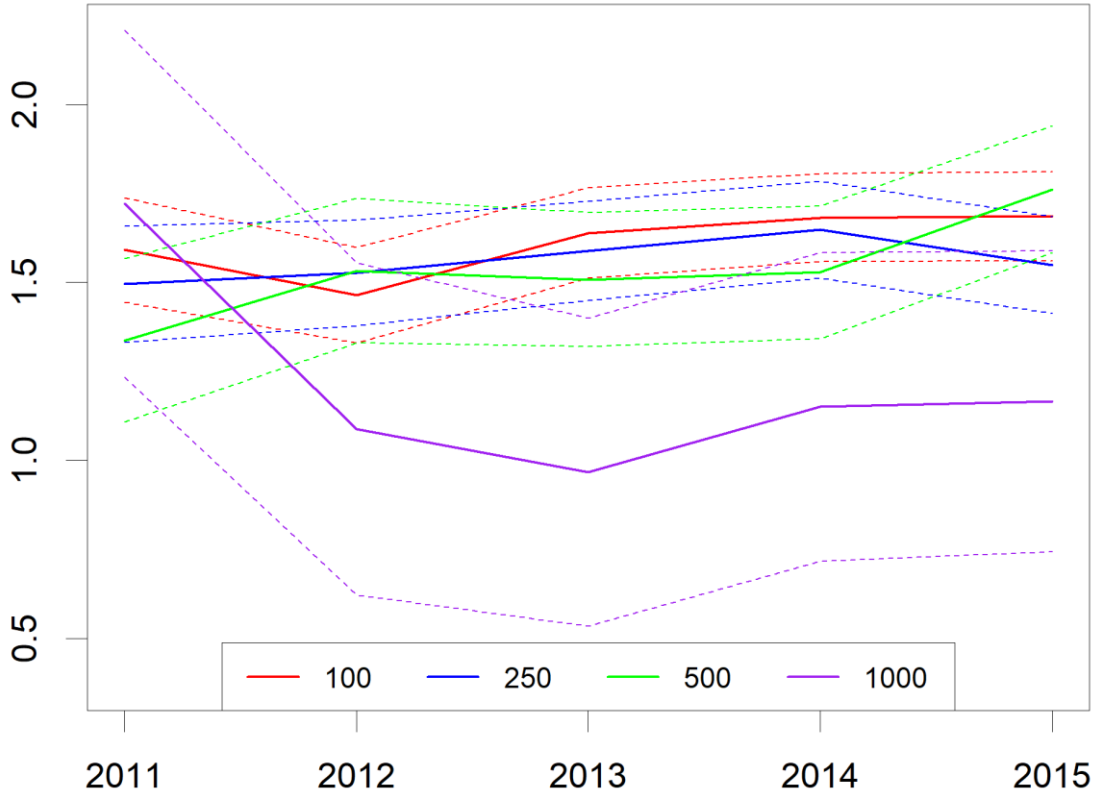
CATLESS2_250

- Impact on 100 relatively similar to CATLESS_100.
- Impact on all other cuts larger than CATLESS2_100.



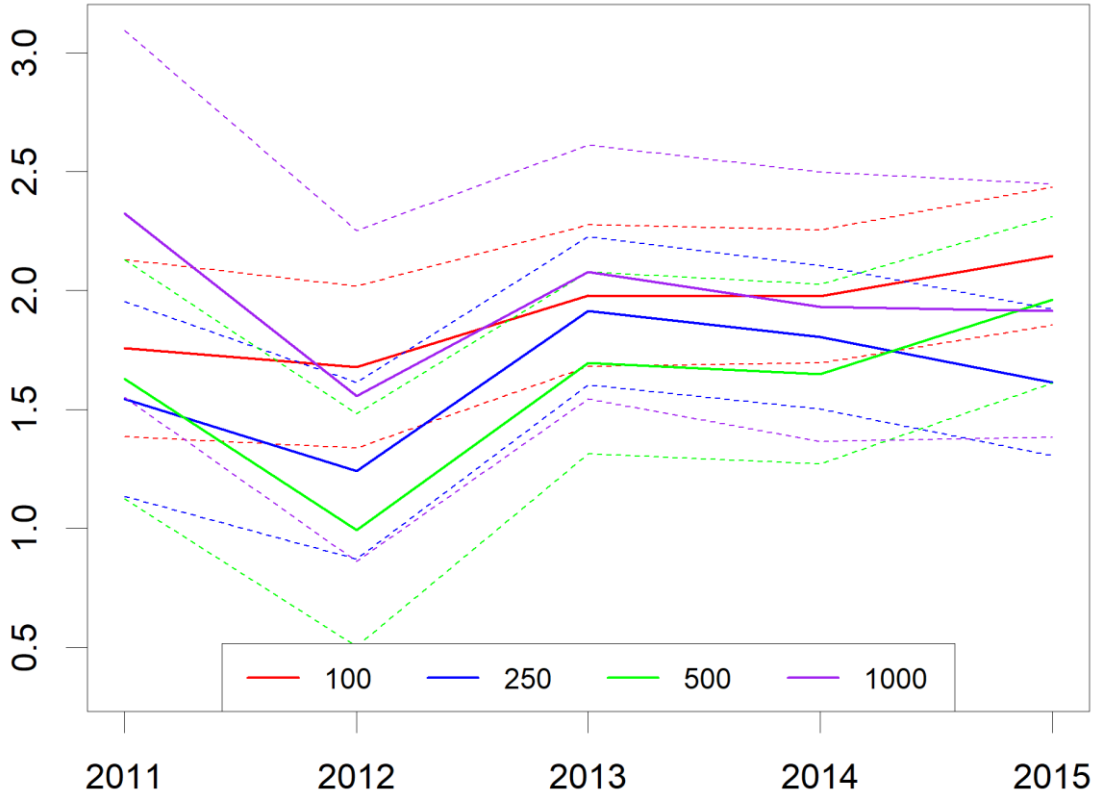
CATLESS2_500

- No significant difference between the various cuts, but all are significantly positive.



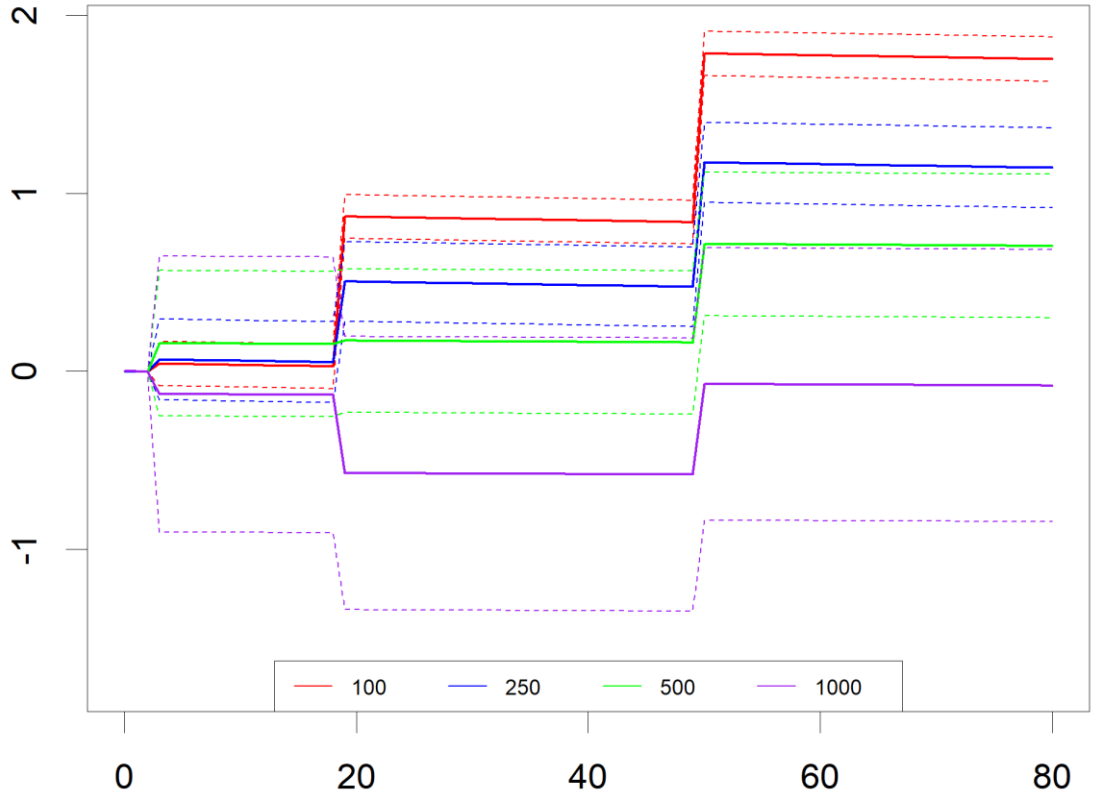
CATLESS2_1000

- Similar to CATLESS2_500, no significant difference between the cuts, but all are significantly positive.



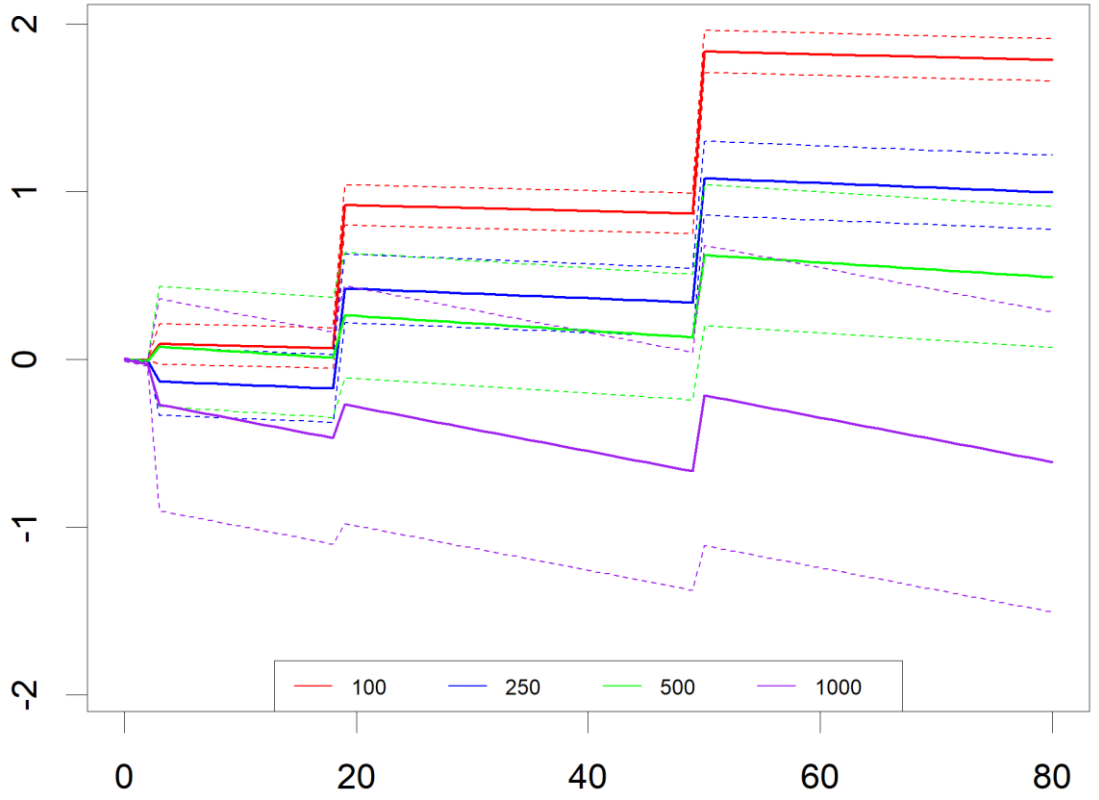
AGE (2011)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000



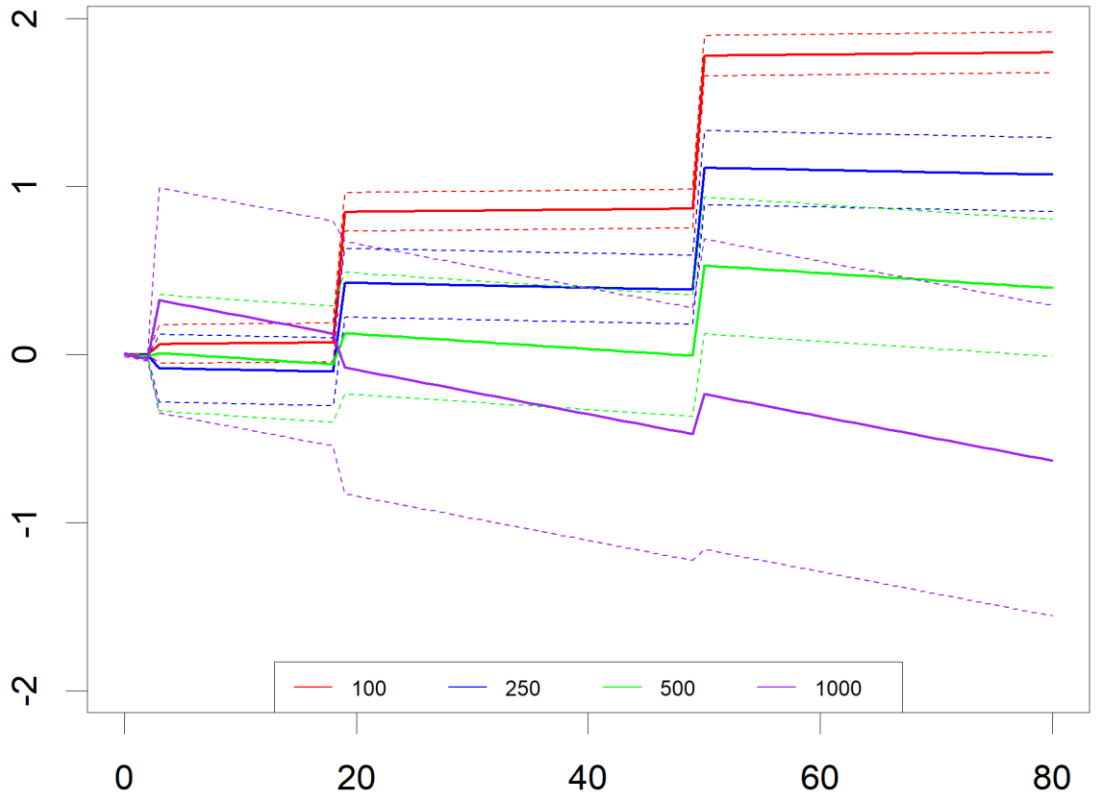
AGE (2012)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000



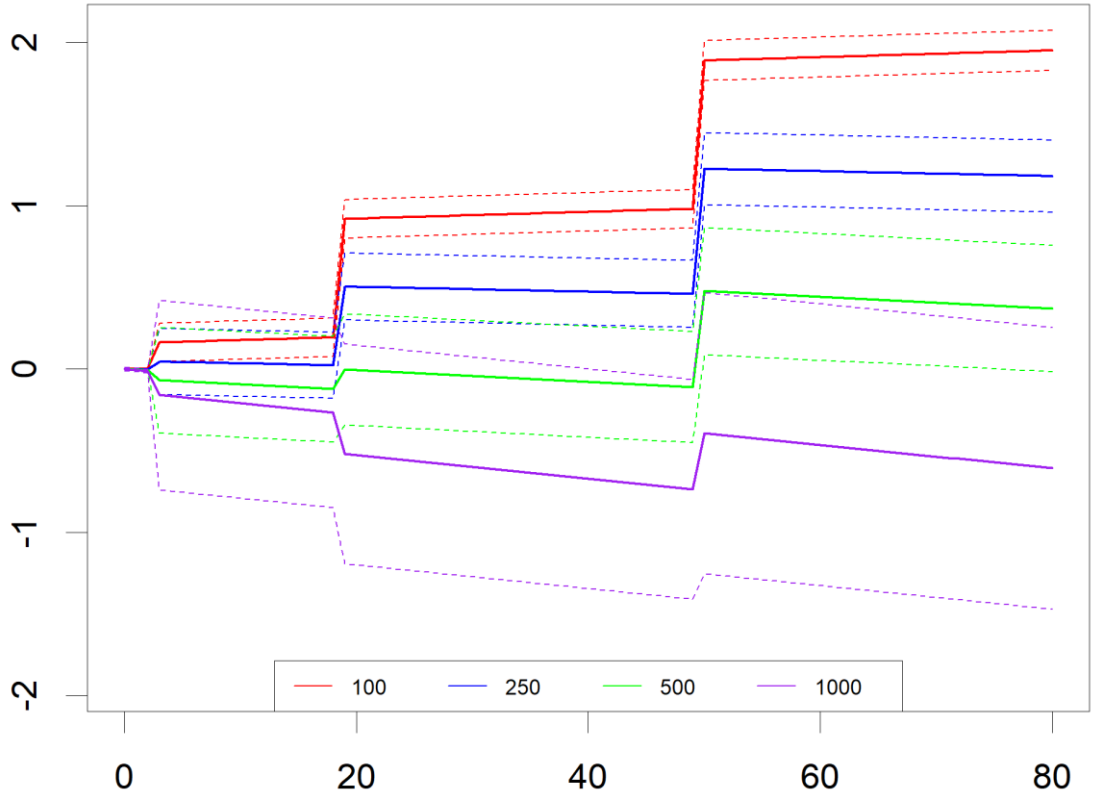
AGE (2013)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000



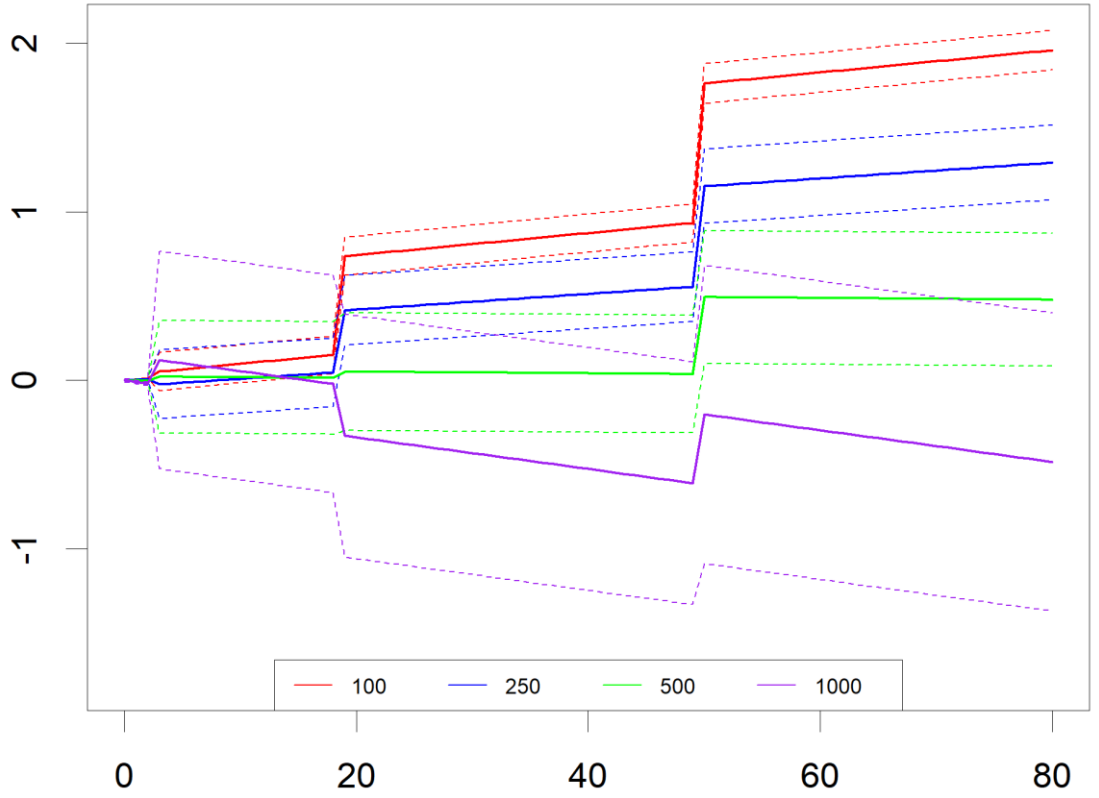
AGE (2014)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000



AGE (2015)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000



Prediction

PREDICTING HIGH-COST MEMBERS IN THE HCCI DATABASE

Prediction

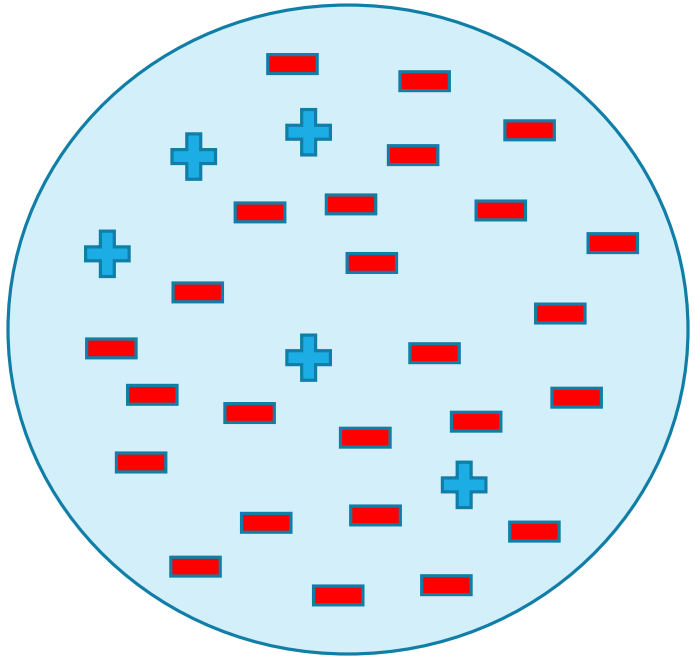
The other main goal of this work is to explore a few possible models for predicting which members are likely to be high-cost next year.

In all cases, we fit the models from training data in one year and use it to predict the following year.

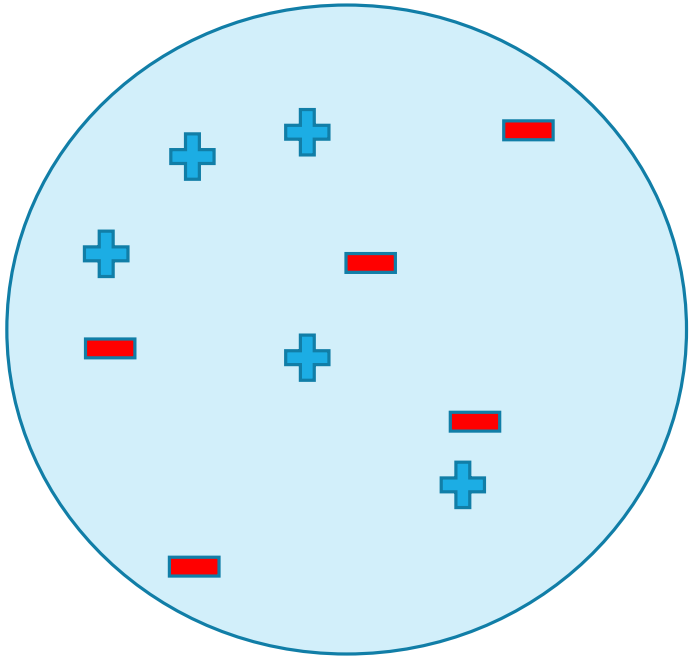
Training Samples

Because predicting an extreme minority class can be very difficult, we compare predictive models based on three different training sets.

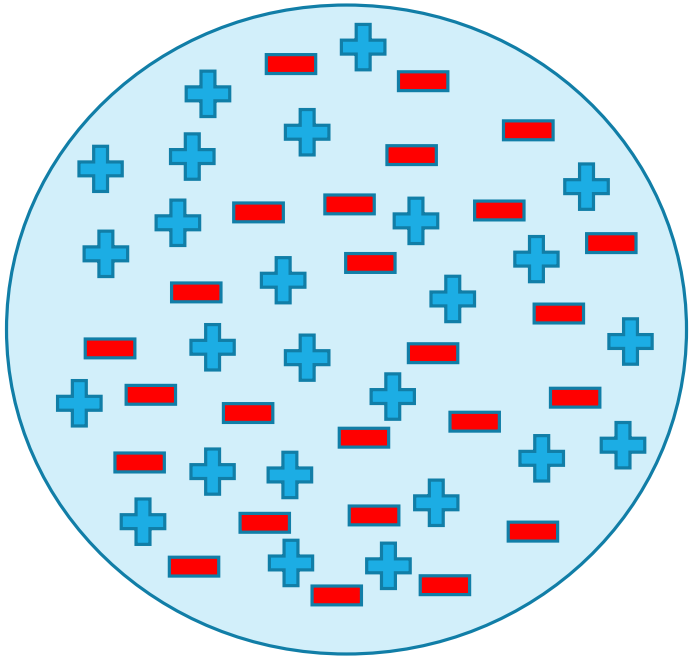
Standard



Undersampled



Oversampled



Methods

To predict which members will be high-cost, we will fit the following models:

- Logistic regression
- Extreme gradient boosted tree (xgboost) using default parameters
- 3 other xgboost models with optimized parameters

Hyperparameters

Maximum tree depth, range (3, 10) - maximum number of branch levels in any tree. A higher number here make 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.

Hyperparameters (continued)

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.

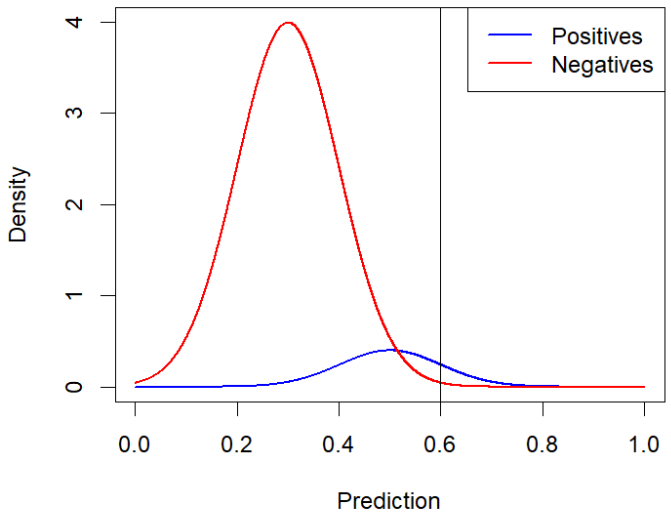
Eta, (0,1) - The learning rate. A higher eta will speed up convergence, while a lower eta may make the convergence more precise.

Learners

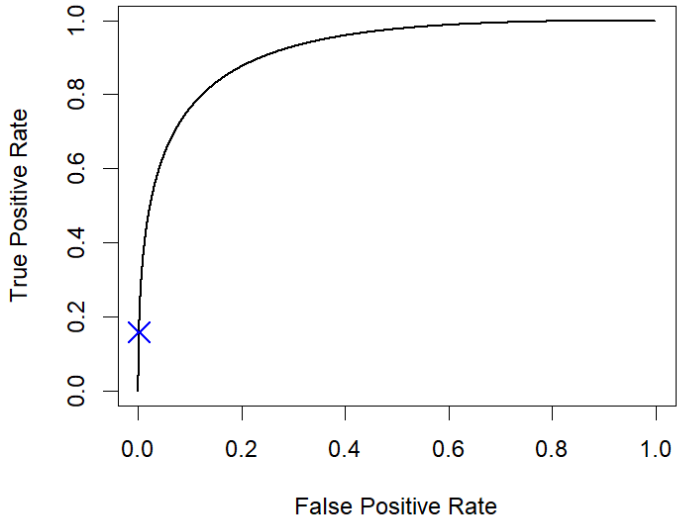
Parameter	Untrained	Trained1	Trained2	Trained3
Maximum Tree Depth	6	3	5	5
Minimum Child Weight	1	9.77	2.98	9.26
Subsample	1	0.66	0.79	0.97
Column Sample by Tree	1	0.76	0.6	0.69
Eta	0.3	0.54	0.52	0.63

Area Under the Curve

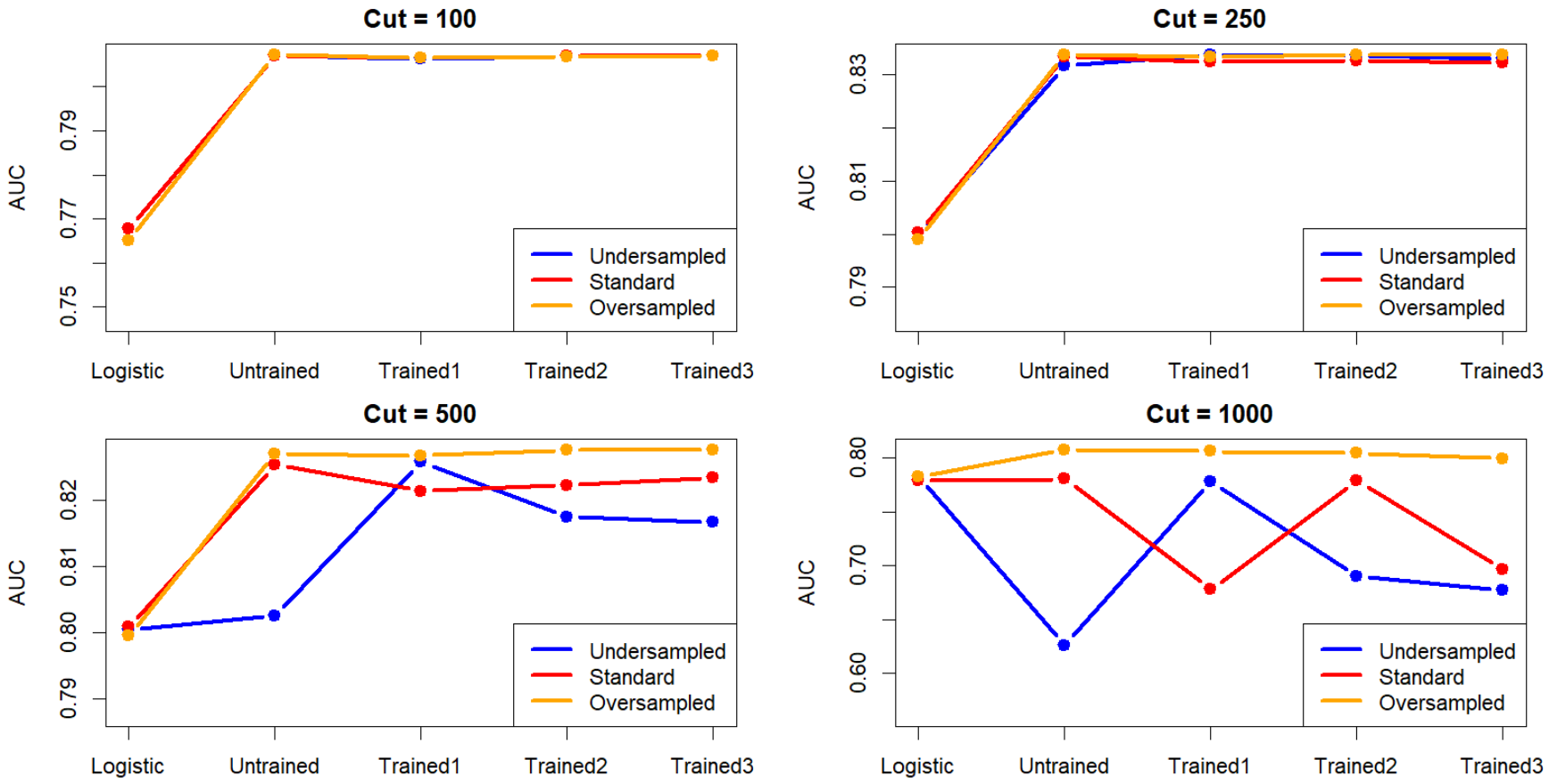
To compare the predictions, we calculate the area under the ROC curve.



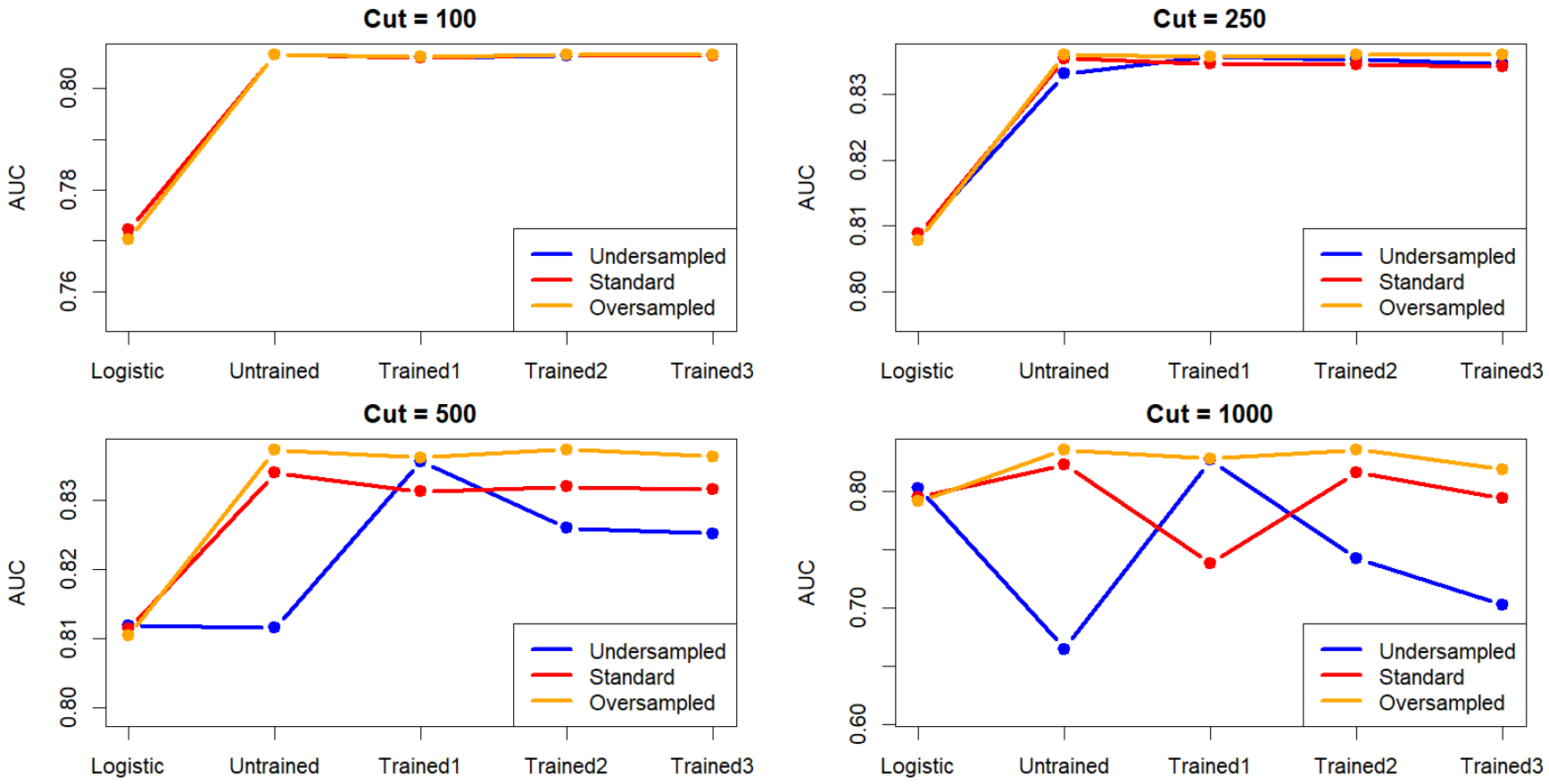
At a threshold of 0.6
True Positive Rate = 0.1587
False Positive Rate = 0.0013



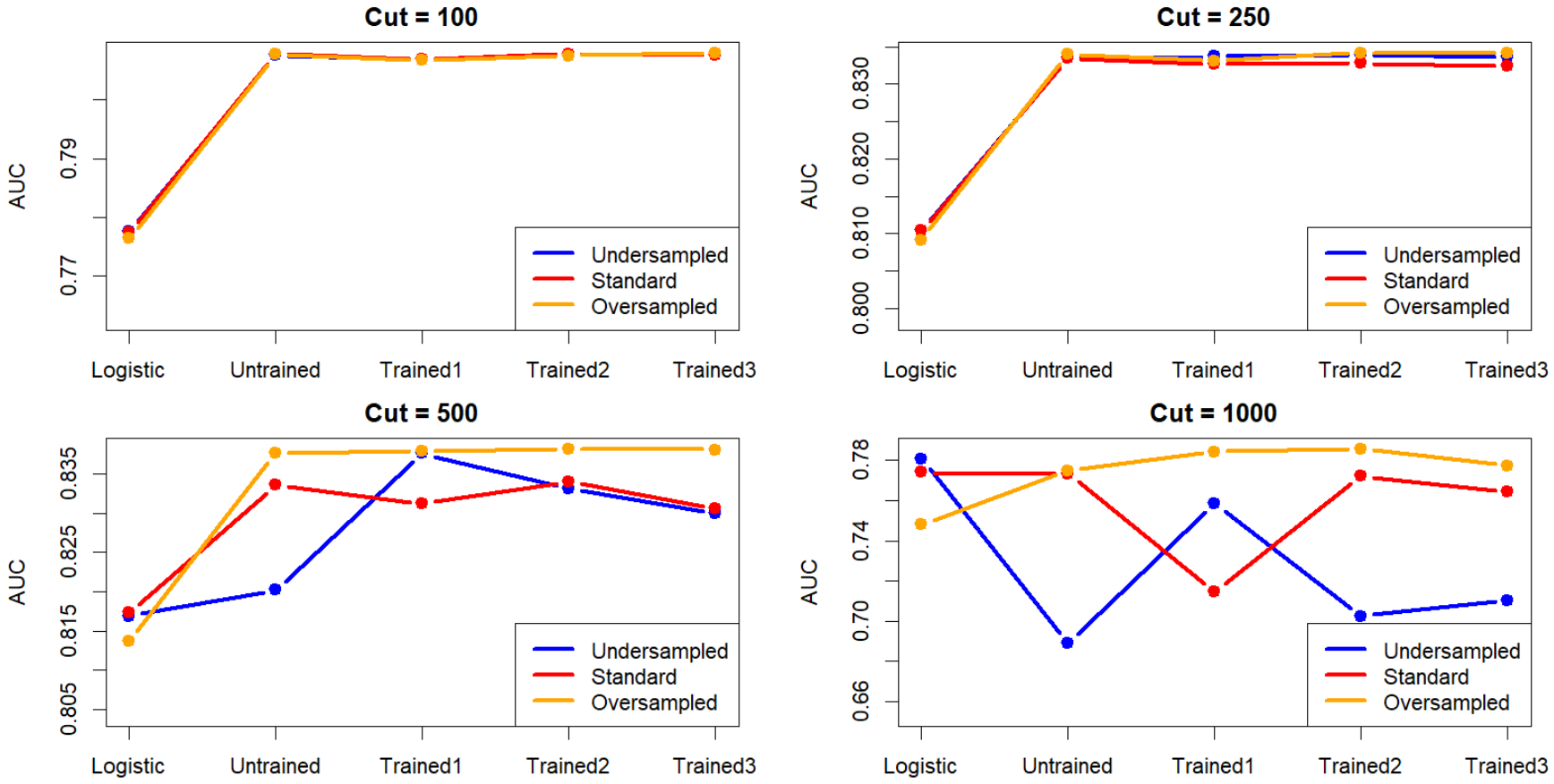
2012


Cut = 100
AUC
0.75
0.77
0.79
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled
Cut = 250
AUC
0.79
0.81
0.83
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled
Cut = 500
AUC
0.79
0.80
0.81
0.82
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled
Cut = 1000
AUC
0.60
0.70
0.80
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled

2013

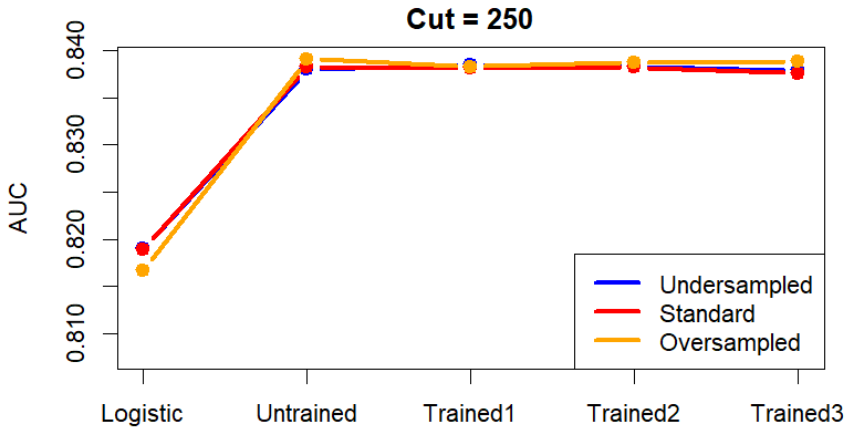

Cut = 100
AUC
0.76
0.78
0.80
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled
Cut = 250
AUC
0.80
0.81
0.82
0.83
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled
Cut = 500
AUC
0.80
0.81
0.82
0.83
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled
Cut = 1000
AUC
0.60
0.70
0.80
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled

2014

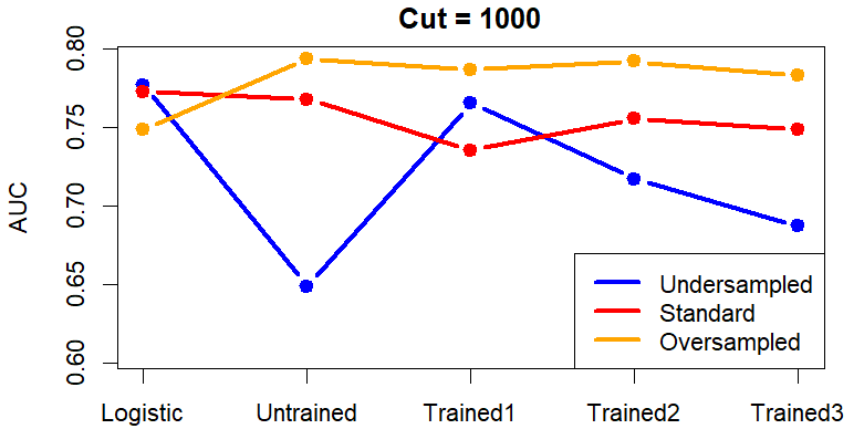

Cut = 100
AUC
0.77
0.79
Undersampled
Standard
Oversampled
Logistic
Untrained
Trained1
Trained2
Trained3
Cut = 250
AUC
0.800
0.810
0.820
0.830
Undersampled
Standard
Oversampled
Logistic
Untrained
Trained1
Trained2
Trained3
Cut = 500
AUC
0.805
0.815
0.825
0.835
Undersampled
Standard
Oversampled
Logistic
Untrained
Trained1
Trained2
Trained3
Cut = 1000
AUC
0.66
0.70
0.74
0.78
Undersampled
Standard
Oversampled
Logistic
Untrained
Trained1
Trained2
Trained3

2015


Cut = 100
AUC
0.77
0.78
0.79
0.80
0.81
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled


Cut = 250
AUC
0.810
0.820
0.830
0.840
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled


Cut = 500
AUC
0.800
0.810
0.820
0.830
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled


Cut = 1000
AUC
0.60
0.65
0.70
0.75
0.80
Logistic
Untrained
Trained1
Trained2
Trained3
Undersampled
Standard
Oversampled

Prediction Notes

Sampling method doesn’t matter too much for cuts 100, and 250 (plenty of positive cases).

For cuts 500 and 1000, oversampling is best.

Undersampled trained1 does almost as well, but trained2 and trained3 do much worse.

Conclusion

While good for inference and understanding the drivers of high-cost members, logistic regression is not the best for prediction.

Oversampling seems to be the best when you have an extreme minority class.

Draft paper available (https://hartman.byu.edu)

This work is threshold-independent, Zoe's work builds on this to incorporate costs.

Predicting High-cost Members in the HCCI Database

BRIAN HARTMAN, BRIGHAM YOUNG UNIVERSITY

JOINT WORK WITH REBECCA OWEN AND ZOE GIBBS