Digital Supplement: Effects of neighborhood-level data on performance equality in predicting 30-day heart failure readmissions at an urban academic medical center

Gary E. Weissman, MD, MSHP, Stephanie Teeple, Nwamaka D. Eneanya, MD, MPH, Rebecca A. Hubbard, PhD, Shreya Kangovi, MD, MSHP

March 3, 2021

Supplemental Methods

Algorithmic Equity

We chose to focus on three measures of algorithmic equity that are both easily measurable and have clear policy remedies.[1–3] First, we examined error due to statistical bias in the predictions. Statistical bias, distinct from human bias, reflects the degree to which, on average, a model's predictions diverge from the true values. Biased models can improve with the collection of additional variables that are informatively associated with the outcome of interest or with the use of a more flexible modeling approach. To measure bias we compared the point estimates of the Brier score between white and non-white patients. Additionally, we iteratively resampled the data with an increasing number of randomly sampled input variables and measured the models' performance on the testing data in aggregate and by patient race.

Second, we examined the error due to variance. This type of error reflects how flexibly the model can generalize to a new dataset (e.g. the testing data) after being fitted with the training dataset. Models that have interpreted random noise in the training data as true signals will not generalize well to new datasets and will be overfit. Error due to variance can be improved with increased regularization or with the collection of more training observations. To measure error due to variance, we iteratively resampled from the training data increasing numbers of observations to train each model and then measured the models' performance using the testing data in aggregate and by patient race.

Third, we examined the classification parity, measured by differences in the positive predictive value, over a range of classification thresholds. In a fair model, these rates should be equivalent across patient sub-groups. This metric of equity is intended to gauge a model that supports deployment of resources for individual patients and does not require that outcome prevalence be the same across subgroups, although this could be an alternative measure of fairness. In predictive model development the classification threshold is often chosen by default, or chosen to maximize a performance metric with little attention to algorithmic equity. Classification parity might be improved in models by increasing the overall model performance in a particular group or by adjusting the classification threshold. Therefore, we examined the positive predictive value in aggregate and by patient race over all possible classification thresholds in the testing sample.

Supplemental Tables

ICD codes to identify congestive heart failure admissions follow the approach of Amarasingham et al.[4]

Table 1: Diagnostic codes for congestive heart failure.

ICD9 Codes
402.01
402.11
402.91
425.1
425.4
425.5
425.7
425.8
425.9
428.0
428.1
428.2
428.21
428.22
428.23
428.3
428.31
428.32
428.33
428.4
428.41
428.42
428.43
428.9

ICD codes for depression are taken from Fiest et al.[5]

Table 2: Diagnostic codes for depression.

ICD9	ICD10
296.20	F32.0
296.21	F32.1
296.22	F32.2

296.23	F32.3
296.24	F32.4
296.25	F32.5
296.30	F32.6
296.31	F32.7
296.32	F32.8
296.33	F32.9
296.34	F33.0
296.35	F33.1
300.4	F33.2
311	F33.3
296.5	F33.8
296.6	F33.9
296.82	F34.1
296.90	F41.2
309.0	F31.3
309.1	F31.4
309.28	F31.5
	F31.6
	F34.8
	F34.9
	F38.0
	F38.1
	F38.8
	F39
	F99

Table 3: Tuning Grid - Elastic Net

alpha	lambda
1e-05	0.001
1e-04	0.003
1e-03	0.005
1e-02	0.010

5e-02	0.015
8e-02	0.020
1e-01	0.025
2e-01	0.030
3e-01	0.040
5e-01	0.050
6e-01	0.100
7e-01	0.150
8e-01	0.200
9e-01	0.300
1e+00	0.400
1e-05	0.500
1e-04	0.700
1e-03	0.800
1e-02	0.900
5e-02	1.000
8e-02	1.500
1e-01	2.000
2e-01	3.000
3e-01	4.000
5e-01	5.000
6e-01	6.000
7e-01	7.000
8e-01	8.000
9e-01	9.000
1e+00	10.000

Table 4: Tuning Grid - Gradient Boosting Machine

n.trees	interaction.depth	shrinkage	n.minobsinnode
5	1	0.0001	1
10	2	0.0010	2
15	3	0.0050	3
20	5	0.0080	4
25	7	0.0100	5
50	10	0.0200	6

75	0.0250	7
100	0.0300	8
	0.0400	9
	0.0500	10
	0.0600	11
	0.0800	12
	0.1000	13
	0.2000	
	0.3000	
	0.4000	
	0.5000	
	0.6000	
	0.7000	
	0.8000	

Table 5: Differences in model performance with inclusion of the Area Deprivation Index with bootstrapped 95% confidence intervals. Positive values indicate an increase in the metric (thus indicating worse performance in this case). Abbreviations: EN = elastic net, GBM = gradient boosting machine, BS = Brier score, CI = confidence interval.

Model type	Metric	Difference	2.5% CI	95% CI	p-value
EN	BS	4.06e-05	-0.0003749	0.0002553	0.7943206
GBM	BS	3.84e-04	-0.0025067	0.0015735	0.7155284

Table 6: Summary of performance characteristics for models across all models in the held-out test set with bootstrapped confidence intervals. Abbreviations: EN = elastic net, GBM = gradient boosting machine.

Model type	ADI	Patient group	Brier score (95% confidence interval)	C-statistic (95% confidence interval)
EN	No	Non-white	0.13 (0.13 to 0.14)	0.60 (0.54 to 0.66)
EN	No	White	0.12 (0.12 to 0.13)	0.64 (0.58 to 0.72)
GBM	No	Non-white	0.14 (0.11 to 0.16)	0.50 (0.44 to 0.56)

GBM	No	White	0.13 (0.10 to 0.16)	0.45 (0.34 to 0.56)
EN	Yes	Non-white	0.13 (0.13 to 0.14)	0.61 (0.56 to 0.66)
EN	Yes	White	0.12 (0.12 to 0.13)	0.64 (0.57 to 0.71)
GBM	Yes	Non-white	0.14 (0.12 to 0.16)	0.40 (0.34 to 0.46)
GBM	Yes	White	0.12 (0.09 to 0.16)	0.42 (0.31 to 0.53)
EN	No	All	0.13 (0.13 to 0.14)	0.60 (0.55 to 0.65)
EN	Yes	All	0.13 (0.13 to 0.14)	0.60 (0.55 to 0.65)
GBM	No	All	0.13 (0.11 to 0.16)	0.48 (0.42 to 0.54)
GBM	Yes	All	0.13 (0.11 to 0.15)	0.40 (0.34 to 0.46)

Table 7: Missingness of predictor variables overall and by race..

var_names	overall_missing_pct	overall_missing_cnt	nonwhite_missingness_cnt	nonwhite_missing_pct	white_missingness_cnt	white_missingness_pct
worst_pco2_24h	0.9460432	1578	0.9598394	1195	0.9054374	383
worst_cpk_24h	0.9322542	1555	0.9236948	1150	0.9574468	405
worst_albumin_24h	0.5941247	991	0.6080321	757	0.5531915	234
worst_bili_24h	0.4862110	811	0.4931727	614	0.4657210	197
worst_troponin_24h	0.4034772	673	0.4248996	529	0.3404255	144
worst_probnp_24h	0.3681055	614	0.3919679	488	0.2978723	126
worst_inr_24h	0.3669065	612	0.3847390	479	0.3144208	133
worst_temp_24h	0.1552758	259	0.1485944	185	0.1749409	74
worst_sbp_24h	0.1522782	254	0.1445783	180	0.1749409	74
worst_wbc_24h	0.0485612	81	0.0514056	64	0.0401891	17
worst_bun_24h	0.0239808	40	0.0240964	30	0.0236407	10
worst_creat_24h	0.0221823	37	0.0232932	29	0.0189125	8

worst_na_24h	0.0209832	35	0.0216867	27	0.0189125	8
worst_glucose_24h	0.0143885	24	0.0136546	17	0.0165485	7
is_hispanic	0.0017986	3	0.0024096	3	0.0000000	0
age	0.0000000	0	0.0000000	0	0.0000000	0
any_cocaine_6mos	0.0000000	0	0.0000000	0	0.0000000	0
any_depr_las6m	0.0000000	0	0.0000000	0	0.0000000	0
any_thc_6mos	0.0000000	0	0.0000000	0	0.0000000	0
count_er_6m	0.0000000	0	0.0000000	0	0.0000000	0
count_h_6m	0.0000000	0	0.0000000	0	0.0000000	0
count_op_6m	0.0000000	0	0.0000000	0	0.0000000	0
gender	0.0000000	0	0.0000000	0	0.0000000	0
has_medicaid	0.0000000	0	0.0000000	0	0.0000000	0
is_white	0.0000000	0	0.0000000	0	0.0000000	0

Table 8: Model performance (Brier score) using an anti-classification approach that removes race entirely from the model and still uses the Area Deprivation Index (ADI) in its place. EN = elastic net, GBM = gradient boosting machine.

Model type	Patient group	Brier Score (95% confidence interval)
EN	All	0.13 (0.13 to 0.14)
GBM	All	0.13 (0.11 to 0.16)
EN	White	0.12 (0.12 to 0.13)
GBM	White	0.13 (0.10 to 0.15)
EN	Non-white	0.13 (0.13 to 0.14)
GBM	Non-white	0.14 (0.11 to 0.16)

Supplemental Figures

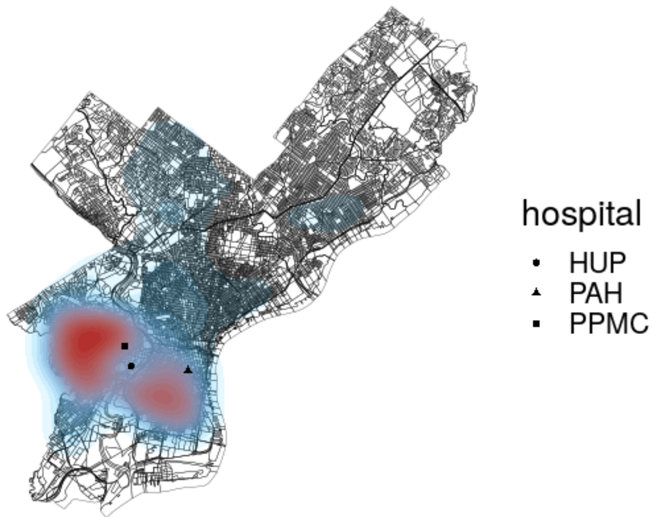

Figure 1: Density plot of patient addresses around Philadelphia with Hospital Locations

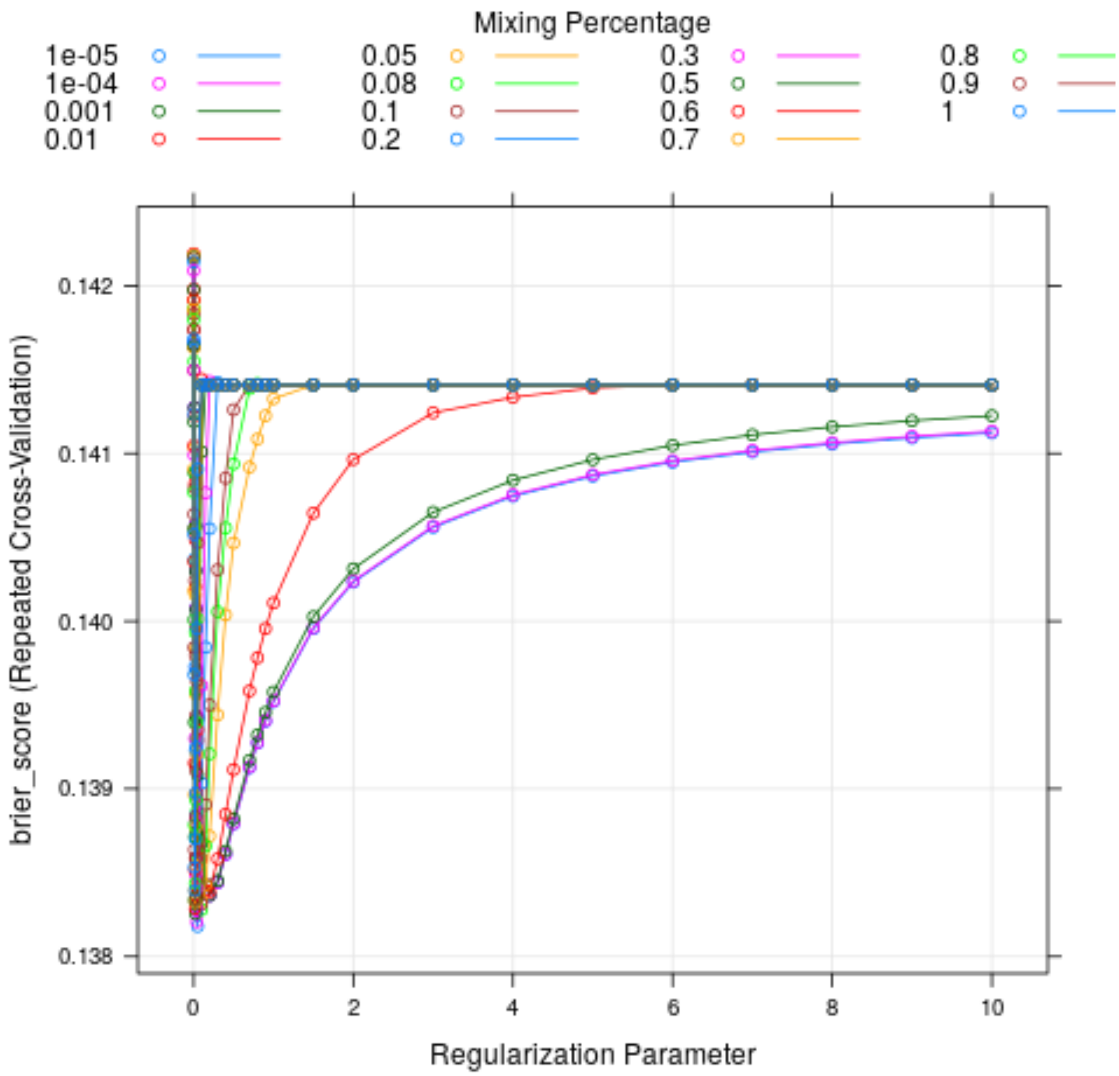

Figure 2: Results of grid search EN with baseline data

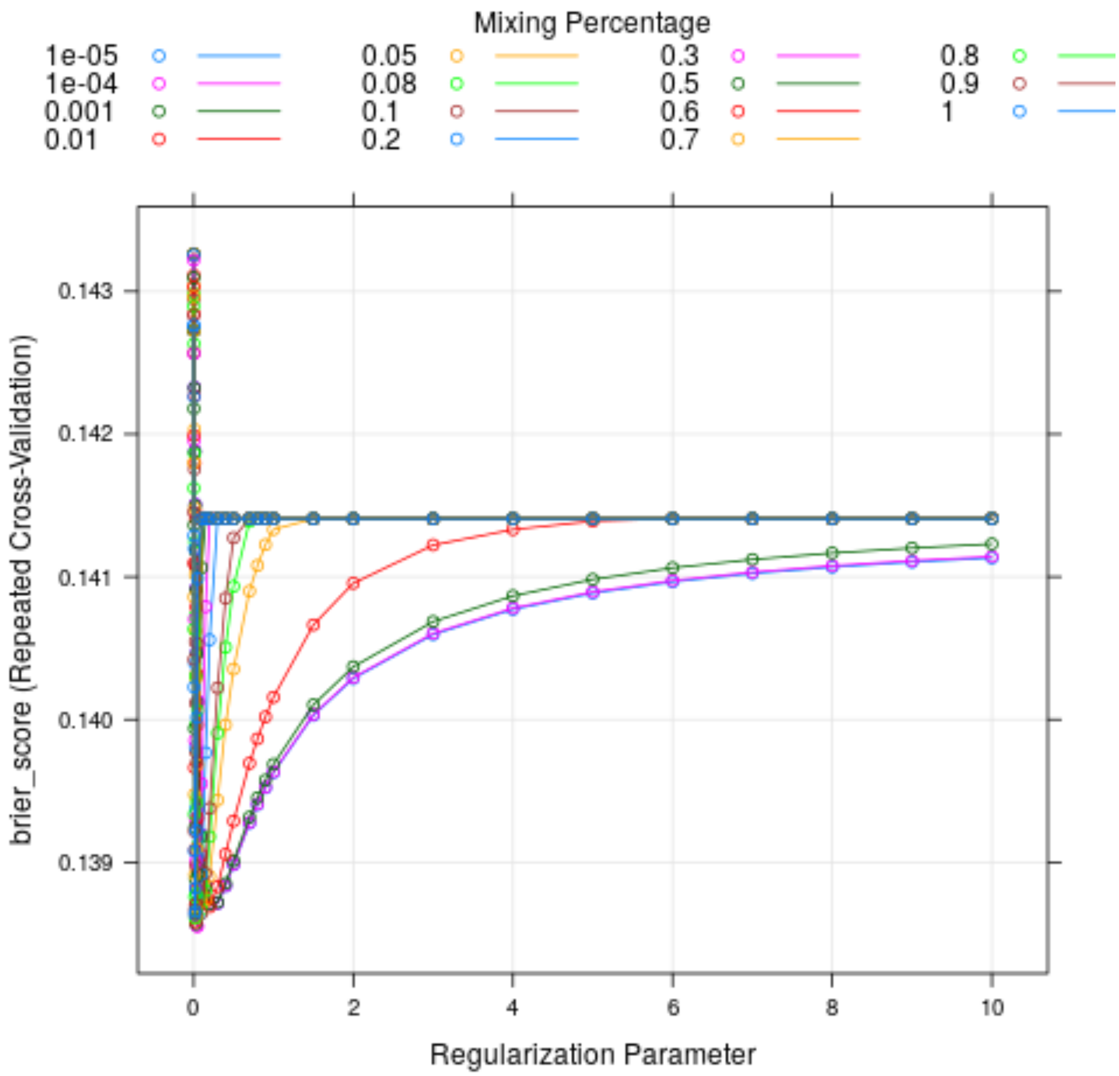

Figure 3: Results of grid search EN with inclusion of ADI data

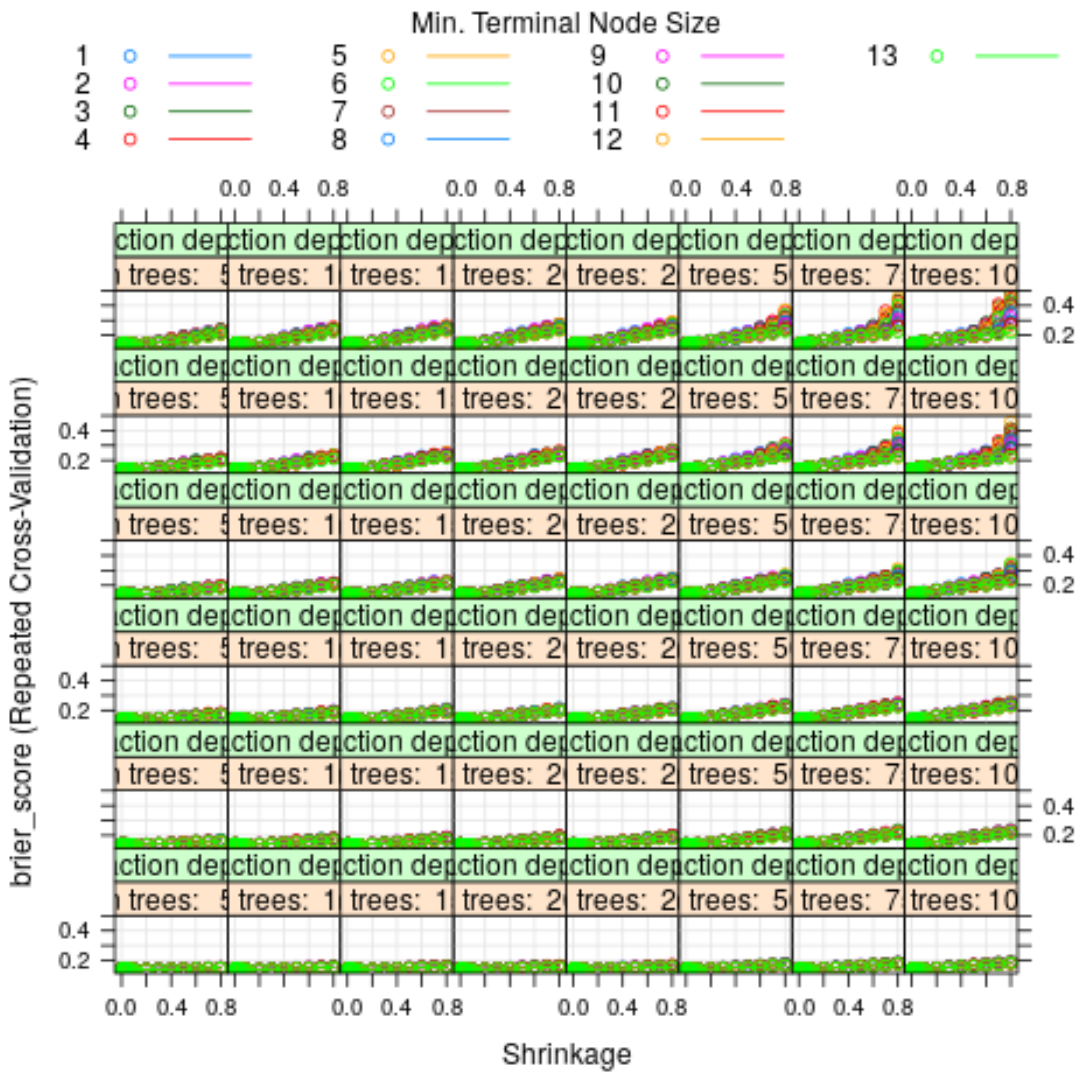

Figure 4: Results of grid search GBM with baseline data

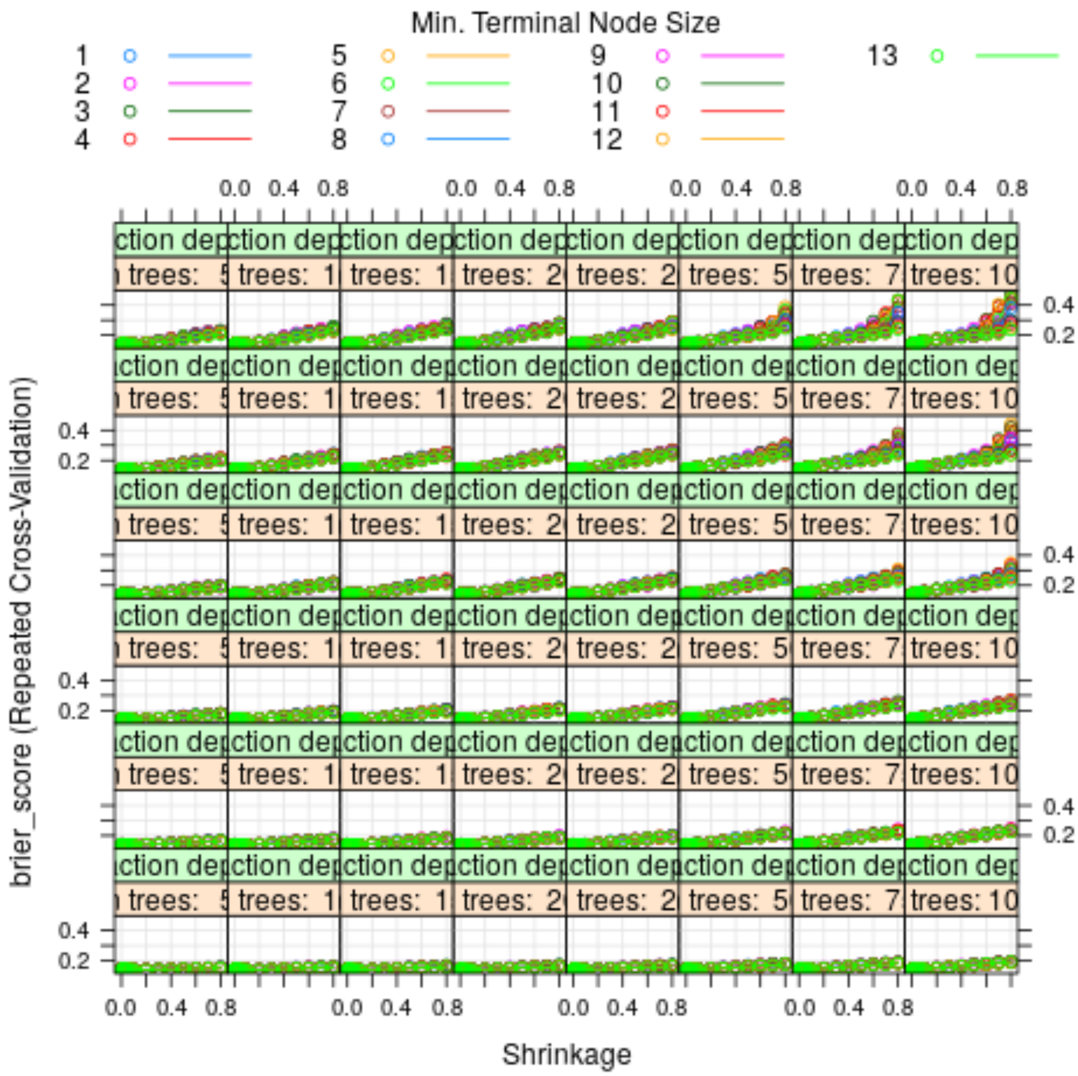

Figure 5: Results of grid search GBM with inclusion of ADI data

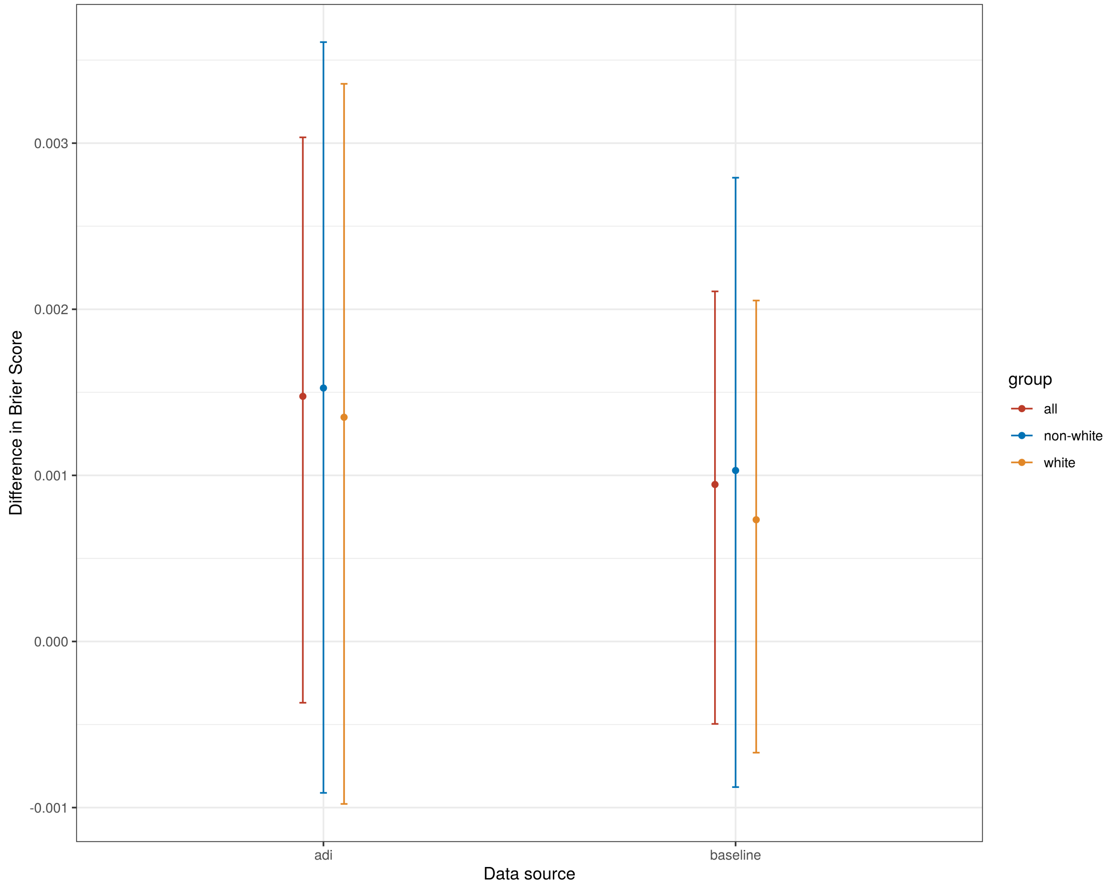

Figure 6: Results of reweighting analysis

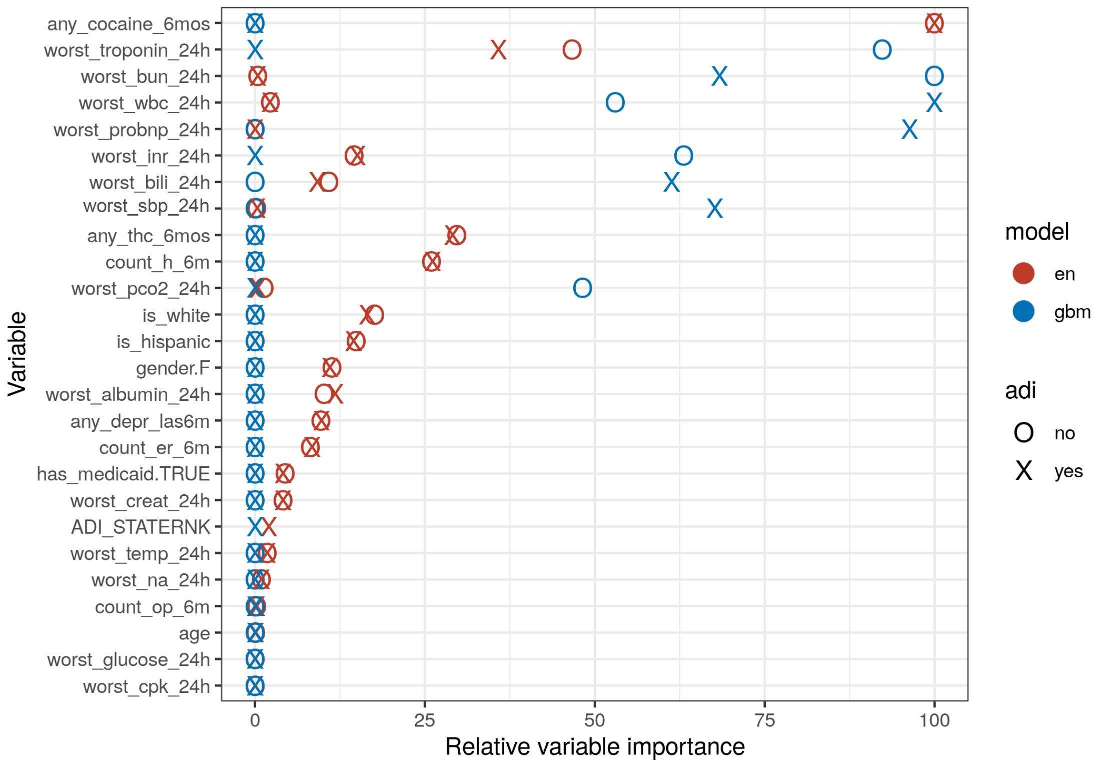

Figure 7: Variable importance by model type and use of the Area Deprivation Index. Multicollinearity between clinical variables such as BUN and creatinine may provide unstable estimates of variable importance for those variables.

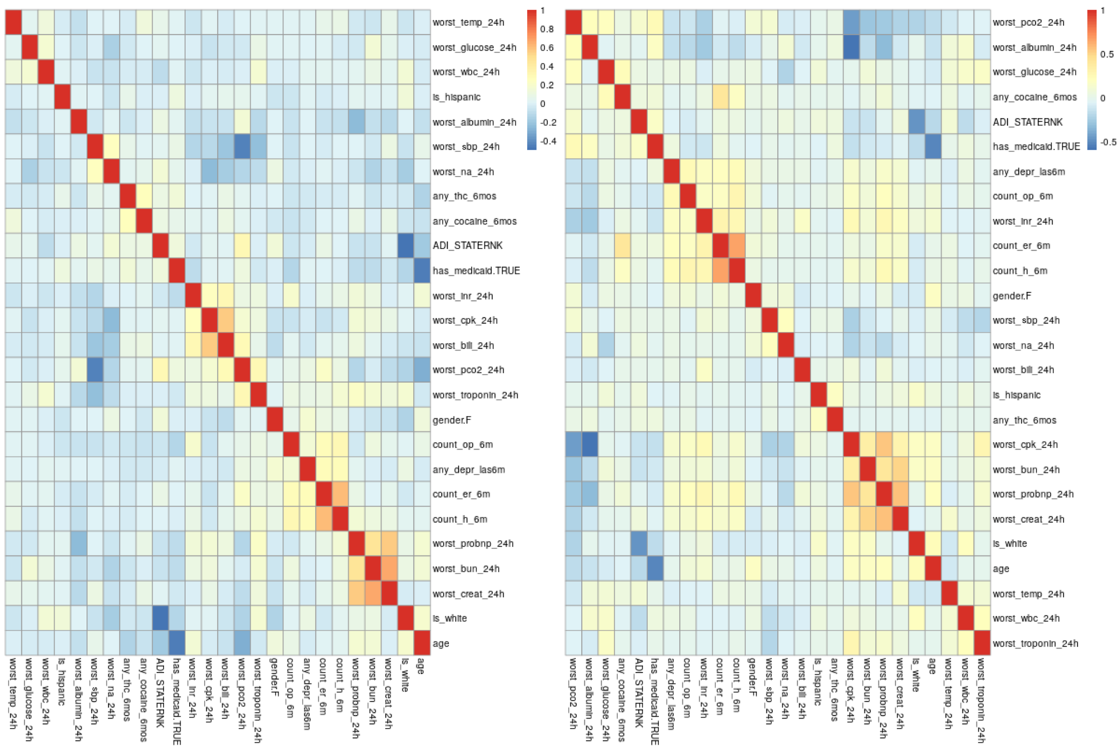

Figure 8: Correlations of predictor variables in the training (left panel) and testing (right panel) sets.

References

1. Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics. 2019 Feb;21(2):167–79.

2. Chen I, Johansson FD, Sontag D. Why Is My Classifier Discriminatory? arXiv:180512002 [cs, stat] [Internet]. 2018 May; Available from: https://arxiv.org/abs/1805.12002

3. Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:180800023 [cs] [Internet]. 2018 Jul; Available from: https://arxiv.org/abs/1808.00023

4. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care. 2010 Nov;48(11):981–8.

5. Fiest KM, Jette N, Quan H, St. Germaine-Smith C, Metcalfe A, Patten SB, et al. Systematic review and assessment of validated case definitions for depression in administrative data. BMC Psychiatry. 2014;14(1):289.