Investigating Medications As An Additional Data Modality In Positive Unlabeled Learning for Predicting Alzheimer's Disease in Electronic Health Records

ABDALLAH FARES[1], Thai Tran[2,3], Mingzhou Fu[2,3], Sriram Sankararaman[4], David A Elashoff [4,5], Keith Vossel[2], Timothy Chang[2]

[1]*BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA* [2]*Department of Neurology, David Geffen School of Medicine, UCLA*
[3]*Medical Informatics Home Area, Department of Bioinformatics, UCLA* [4]*Computational Medicine, UCLA* [5]*Department of Biostatistics, UCLA*

UCLA Neurology

UCLA College | Life Sciences Computational & Systems Biology

BACKGROUND

- Alzheimer's Disease (AD) is underdiagnosed, particularly in underrepresented racial and ethnic groups
- Prior AD prediction studies focused on diverse groups:
  - Rely on expensive labeled data
  - Rarely address racial disparities in model performance
- Previously we proposed a semi-supervised PUL (SSPUL) framework, which couples PUL with pre- and post-processing bias mitigation approaches on diverse EHR data to accurately predict undiagnosed AD among diverse groups
- This prior framework, however, relied exclusively on demographics and diagnostic data as predictors, limiting the feature set and potentially model performance
- **Here we extend the SSPUL framework to incorporate medication data alongside diagnostics, leveraging elastic net feature selection to mitigate the collinearity between the two (as diagnoses determine medications)**

METHODS

Figure 1. Original study design

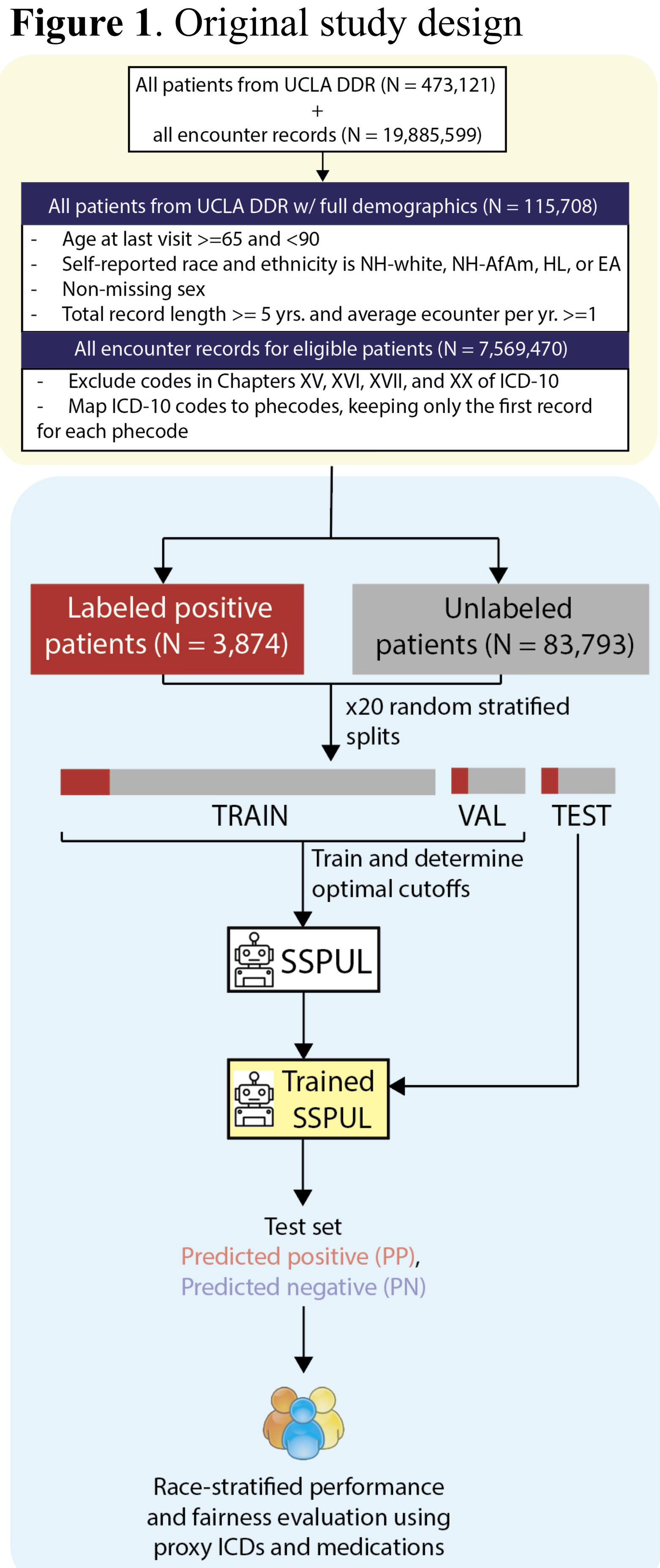


Figure 2. Feature selection breakdown

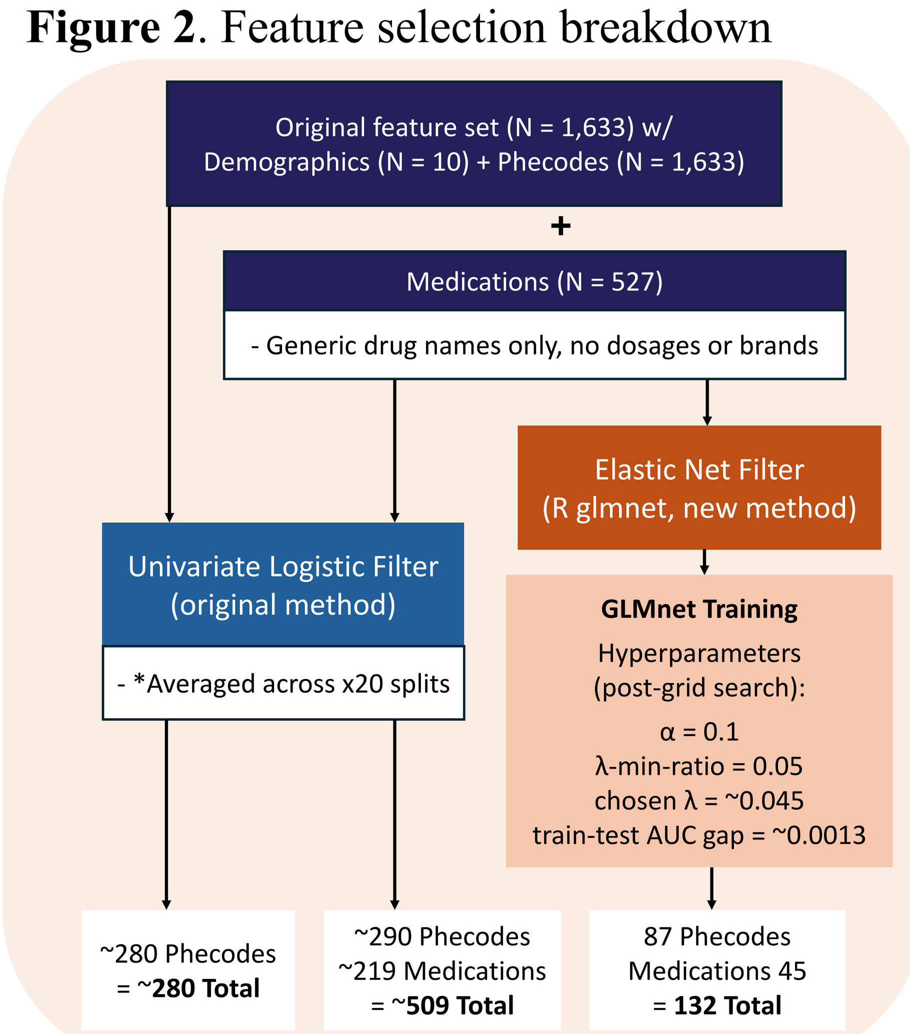
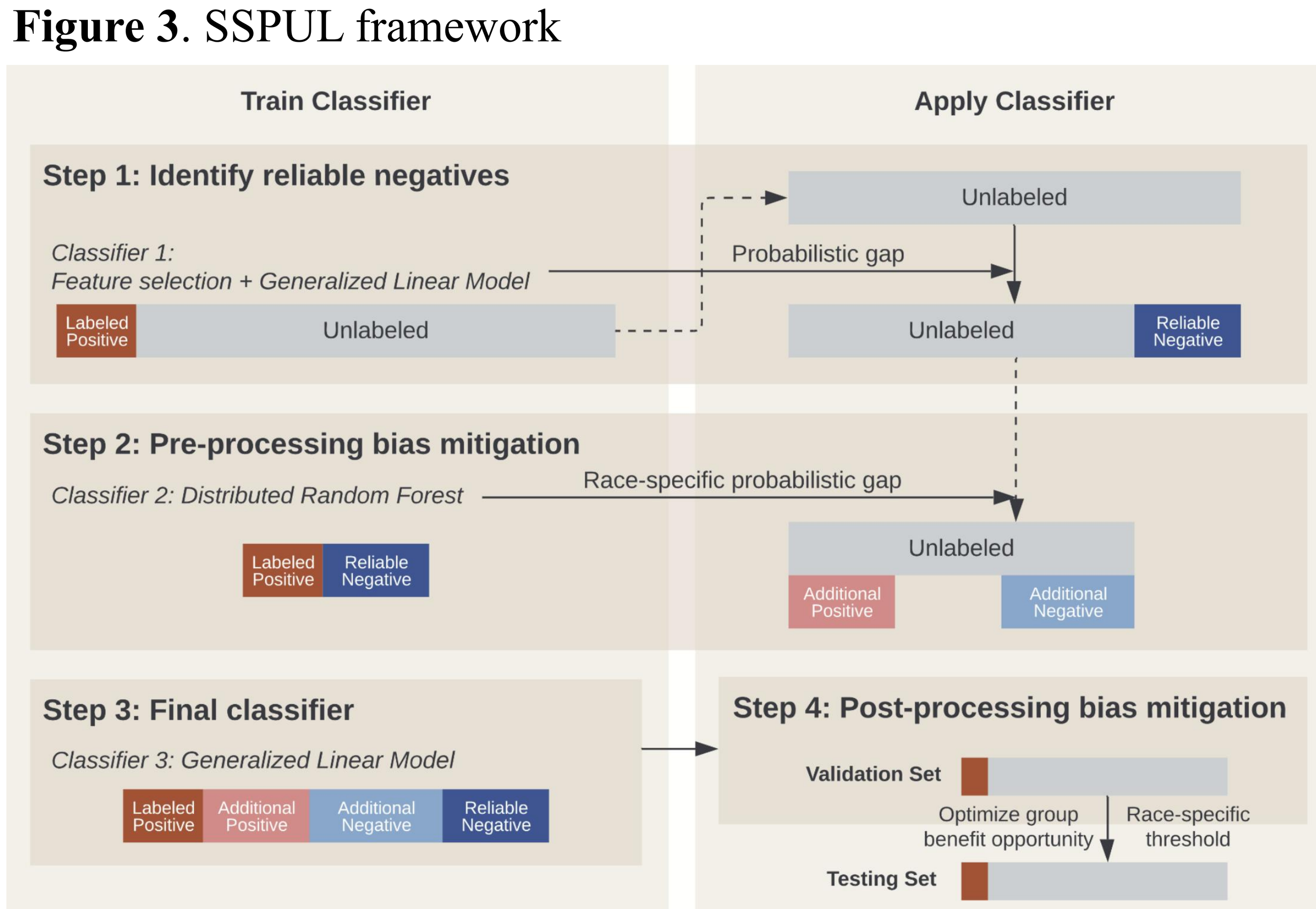


Figure 3. SSPUL framework



RESULTS

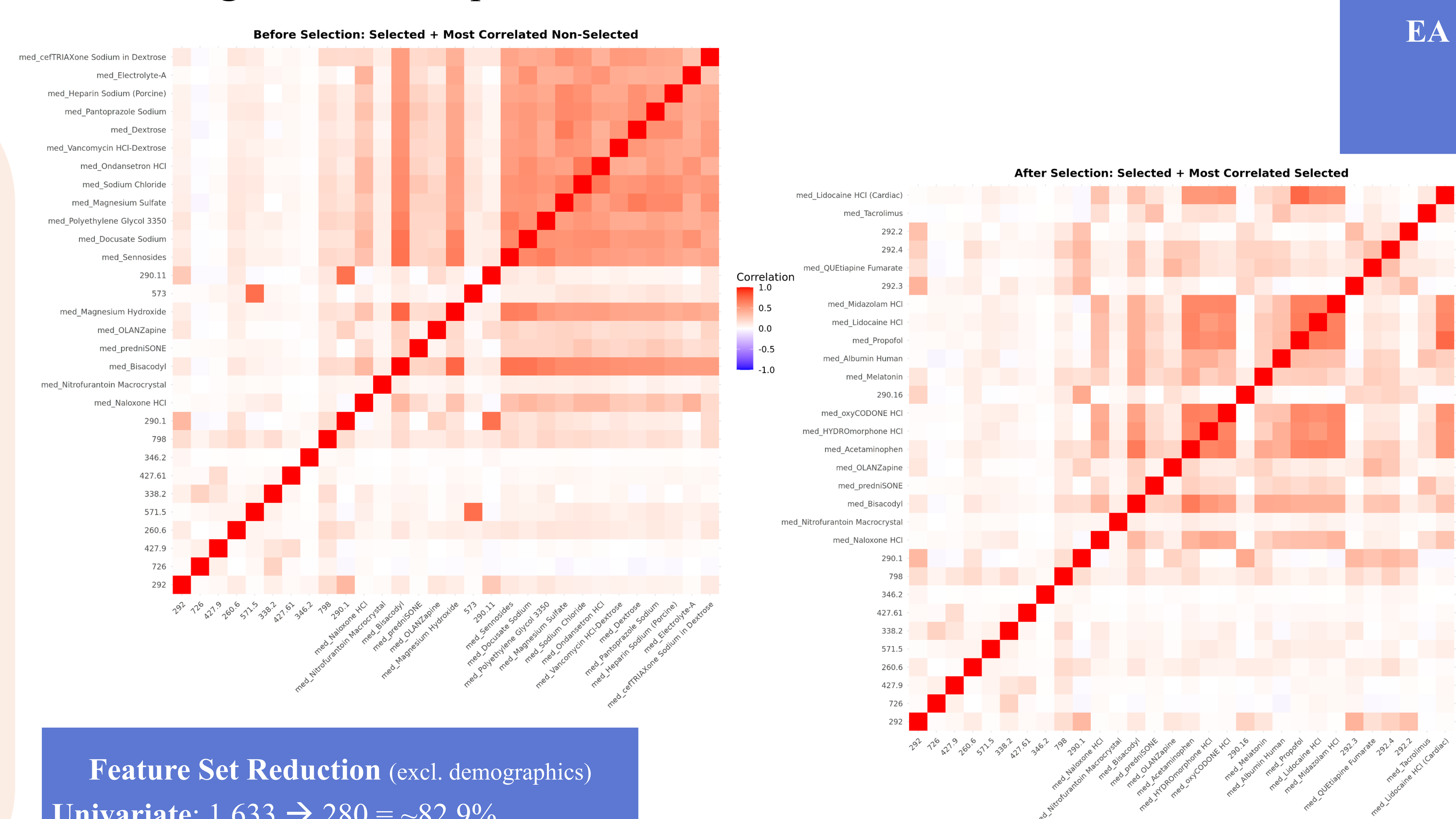
Table 1. Test set performance of SSPUL (GBE) and baseline models

Feature Selection Method	Model	Sensitivity	Precision	Specificity	AUCPR
Univariate Logistic (diagnostics only) (previous model method and results)	Supervised (risk factors/MCC)	0.453 ± 0.083	0.299 ± 0.028	0.804 ± 0.054	0.305 ± 0.011
	Supervised (full/MCC)	0.458 ± 0.029	**0.848 ± 0.035**	**0.985 ± 0.005**	0.673 ± 0.014
	SSPUL (GBE)	**0.793 ± 0.038**	0.789 ± 0.037	0.962 ± 0.007	**0.852 ± 0.026**
Univariate Logistic (diagnostics + medications)	Supervised (risk factors/MCC)	0.473 ± 0.033	0.295 ± 0.009	0.795 ± 0.022	0.310 ± 0.011
	Supervised (full/MCC)	0.451 ± 0.041	**0.812 ± 0.042**	**0.981 ± 0.006**	0.661 ± 0.01
	SSPUL (GBE)	**0.764 ± 0.033**	0.746 ± 0.01	0.953 ± 0.002	**0.827 ± 0.015**
1-SE Max GLMnet (diagnostics + medications)	Supervised (risk factors/MCC)	0.473 ± 0.033	0.295 ± 0.009	0.795 ± 0.022	0.310 ± 0.011
	Supervised (full/MCC)	0.455 ± 0.039	**0.834 ± 0.037**	**0.983 ± 0.005**	0.670 ± 0.017
	SSPUL (GBE)	**0.808 ± 0.029**	0.795 ± 0.032	0.962 ± 0.006	**0.861 ± 0.019**

Table 2. Test set performance of SSPUL (GBE) and baseline models by race/ethnicity

Race / Ethnicity	Model	Sensitivity	Precision	Specificity	AUCPR
NH-white	Supervised (risk factors/MCC)	0.473 ± 0.041	0.286 ± 0.01	0.79 ± 0.016	0.300 ± 0.018
	Supervised (full/MCC)	0.455 ± 0.041	**0.864 ± 0.035**	**0.987 ± 0.005**	0.677 ± 0.017
	SSPUL (GBE)	**0.812 ± 0.036**	0.804 ± 0.033	0.965 ± 0.006	**0.872 ± 0.019**
NH-AfAm	Supervised (risk factors/MCC)	0.554 ± 0.025	0.306 ± 0.052	0.748 ± 0.037	0.373 ± 0.039
	Supervised (full/MCC)	0.512 ± 0.051	**0.818 ± 0.052**	**0.978 ± 0.008**	0.693 ± 0.03
	SSPUL (GBE)	**0.816 ± 0.059**	0.806 ± 0.061	0.961 ± 0.014	**0.891 ± 0.03**
HL	Supervised (risk factors/MCC)	0.458 ± 0.071	0.286 ± 0.035	0.802 ± 0.029	0.293 ± 0.044
	Supervised (full/MCC)	0.535 ± 0.059	0.672 ± 0.067	0.954 ± 0.017	0.652 ± 0.043
	SSPUL (GBE)	**0.815 ± 0.056**	**0.787 ± 0.054**	**0.962 ± 0.01**	**0.871 ± 0.031**
EA	Supervised (risk factors/MCC)	0.452 ± 0.003	0.328 ± 0.012	0.821 ± 0.023	0.347 ± 0.031
	Supervised (full/MCC)	0.396 ± 0.049	**0.869 ± 0.055**	**0.988 ± 0.006**	0.661 ± 0.01
	SSPUL (GBE)	**0.790 ± 0.019**	0.765 ± 0.045	0.952 ± 0.01	**0.823 ± 0.025**

Figure 4. Heatmaps of 15 feature correlations before/after selection



Feature Set Reduction (excl. demographics)
Univariate: 1,633 → 280 = ~82.9%
Univariate w/ Meds: 2160 → 509 = ~76.4%
GLMnet w/ Meds: 2160 → 132 = ~93.8%

Figure 5. Evaluation of fairness across models

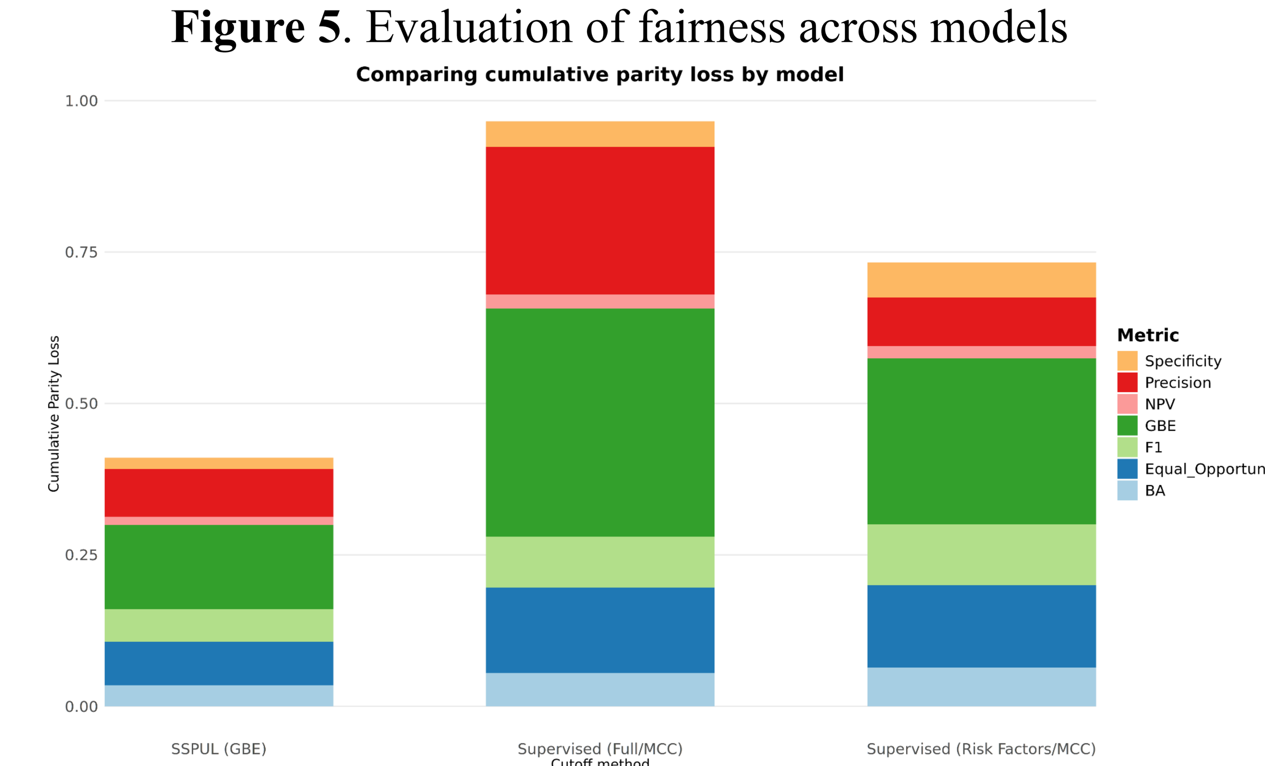


Table 3. Top 10 1-SE Max GLMnet selected medications

Medication	Rank	Coefficient	Description
Quetiapine Fumarate	13	0.2052	Antipsychotic
Alteplase	15	-0.1811	Plasminogen activator
Citalopram Hydrobromide	20	0.1323	Antidepressant
Montelukast Sodium	30	-0.0879	Asthma treatment
Tacrolimus	36	-0.0710	Immunosuppressant
Influenza Vaccine	38	0.0677	Vaccine
Sertraline HCl	41	0.0621	Antidepressant
Olanzapine	42	0.0619	Antipsychotic
Mannitol	43	-0.0615	Asthma treatment
Dexamethasone	45	0.2801	Immunosuppressant

CONCLUSIONS

- Using elastic net for feature selection successfully reduced the feature set to still produce statistically similar results
- Medications and diagnoses together as features (likely due to their collinearity) do not significantly aid AD detection in this setting
- **Elastic net feature selection could prove useful as an addition to the SSPUL pipeline, especially as more data modalities are incorporated**

FUTURE DIRECTION

- Incorporating genetic and temporal data (far more separate features in terms of correlation)
- Perform validation using chart review (gold standard)
- Optimize GBE with respect to both race and ethnicity and sex