

Leveraging APHRODITE to identify bias in statistical phenotyping algorithms

PRESENTER: **Juan M. Banda**

INTRO:

- The widespread adoption of machine learning (ML) algorithms for risk-stratification has resulted in documented cases of racial/ethnic biases within algorithms (1,2). When built without careful weightage and bias-proofing, ML algorithms can give recommendations which worsen health disparities faced by communities of color (3).
- Systematic differences in the output of statistical phenotyping algorithms for vulnerable populations is largely unexplored, particularly within the Observational Health Sciences and Informatics (OHDSI) community and tools.
- By leveraging APHRODITE (4,5), a probabilistic phenotyping framework, we examine four clinical conditions -- **dementia, frailty, mild cognitive impairment and Alzheimer's disease** -- common in vulnerable older adults. We aim to automate the process of identifying the presence of bias in phenotyping algorithms, by providing a standard and automatic framework for their assessment.

Our initial evaluation elucidates that the selected phenotype algorithms have performance (precision, recall, accuracy) variations anywhere between 3% to 30% across ethnic populations; even when not using ethnicity as a feature. Demonstrating how important it is to assess these models' performance for specific subgroups before deploying them in routine use



Scan for more info

A) Alzheimer's disease						B) Frailty					
	Asian	White	Black	Native American	Pacific Islander		Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	1.01%	1.78%	5.02%	2.92%	Asian Model	0.00%	19.30%	16.86%	6.52%	11.57%
White Model	0.14%	0.00%	0.99%	5.84%	0.49%	White Model	17.31%	0.00%	17.72%	3.36%	1.57%
Black Model	1.85%	1.51%	0.00%	0.50%	2.86%	Black Model	4.15%	9.23%	0.00%	17.44%	10.46%
Native A. Model	4.32%	4.66%	2.51%	0.00%	2.33%	Native A. Model	20.32%	4.41%	19.90%	0.00%	14.38%
Pacific I. Model	0.62%	0.25%	2.57%	8.34%	0.00%	Pacific I. Model	15.74%	17.70%	5.35%	6.85%	0.00%

C) Mild Cognitive Impairment						D) Dementia					
	Asian	White	Black	Native American	Pacific Islander		Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	7.70%	0.24%	4.29%	4.73%	Asian Model	0.00%	0.02%	0.63%	0.91%	0.71%
White Model	20.71%	0.00%	18.28%	6.66%	15.32%	White Model	0.42%	0.00%	1.58%	1.26%	1.22%
Black Model	18.20%	13.47%	0.00%	4.69%	18.12%	Black Model	0.85%	0.74%	0.00%	3.00%	2.21%
Native A. Model	4.47%	24.82%	3.26%	0.00%	18.13%	Native A. Model	0.46%	0.07%	0.21%	0.00%	4.64%
Pacific I. Model	2.42%	20.68%	7.20%	22.08%	0.00%	Pacific I. Model	4.20%	4.85%	3.11%	5.94%	0.00%

Figure 1. Variation of classification accuracy for the Random Forest models across phenotypes

Methods

- We created an experimental framework, on top of APHRODITE, to explore racial/ethnic biases within a single healthcare system, Stanford Health Care, to fully evaluate the performance of such algorithms under different ethnicity distributions
- Cases and controls are matched by age/gender/race/length of record. We evaluated three different classification algorithms (LASSO, Random Forest, and Support Vector Machines). Each of the standard concepts in the OMOP CDM corresponding to racial categories are evaluated.
- We built models using a single race group and tested them against all other race groups (Figure 1). In the full study we additionally used seven different evaluations: traditional model (all data available), balanced model (per race), leave-one-out combinations. For all models we used 10-fold cross validation. Note that we removed all patients with an Unknown race from this evaluation.

This work is supported by the National Institute of Aging through Stanford University's Stanford Aging & Ethnogeriatrics Transdisciplinary Collaborative Center (SAGE) center (award 3P30AG059307-02S1)

Juan M. Banda¹, Nigam H. Shah²,
Vyjeyanthi S Periyakoil²
¹Georgia State University,
Atlanta, GA, USA
²Stanford University
School of Medicine
Stanford, CA, USA

