Leveraging APHRODITE to identify bias in statistical phenotyping algorithms

Juan M. Banda, Nigam H. Shah, Vyjeyanthi S Periyakoil

Background

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.

Methods

For this work, 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. Doing so will allow us to identify which algorithms may be biased and under what conditions. In total we have four phenotypes, each with more than a few thousand patients (cases) as shown in Table 1, the 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. To investigate the presence of racial bias in the models, 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.

Phenotype	Cases	Controls	Gender		Race					
			Female	Male	Asian	Black	Native A.	Pacific Islander	White	Unk.
Dementia	16,998	16,998	56.22%	43.78%	11.07%	4.96%	0.27%	0.80%	60.33%	22.57%
Frailty	15,133	15,133	55.41%	44.59%	14.72%	5.45%	0.38%	1.36%	55.58%	22.51%
Mild Cognitive Impairment	8,292	8,292	49.82%	50.18%	11.73%	3.88%	0.22%	0.92%	60.06%	23.19%
Alzheimer's disease	12,828	12,828	60.05%	39.95%	12.08%	4.85%	0.25%	0.69%	63.03%	19.11%

Table 1. Demographic distribution for our phenotype cases and controls.

Results

Having more than 144 models to evaluate on seven datasets, we will be brief in this section, presenting

highlevel highlights of our findings. In order to properly describe and evaluate our models, we used the following metrics: **accuracy** - fraction of assignments the model identified correctly, **sensitivity** - proportion of positives that are correctly identified, and **specificity** - proportion of negatives that are correctly identified. For simplicity of this submission, we will only present results for the accuracy metric in Figure 1.

A) Alzheimer's disease

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	1.01%	1.78%	5.02%	2.92%
White Model	0.14%	0.00%	0.99%	5.84%	0.49%
Black Model	1.85%	1.51%	0.00%	0.50%	2.86%
Native A. Model	4.32%	4.66%	2.51%	0.00%	2.33%
Pacific I. Model	0.62%	0.25%	2.57%	8.34%	0.00%

B) Frailty

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	19.30%	16.86%	6.52%	11.57%
White Model	17.31%	0.00%	17.72%	3.36%	1.57%
Black Model	4.15%	9.23%	0.00%	17.44%	10.46%
Native A. Model	20.32%	4.41%	19.90%	0.00%	14.38%
Pacific I. Model	15.74%	17.70%	5.35%	6.85%	0.00%

C) Mild Cognitive Impairment

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	7.70%	0.24%	4.29%	4.73%
White Model	20.71%	0.00%	18.28%	6.66%	15.32%
Black Model	18.20%	13.47%	0.00%	4.69%	18.12%
Native A. Model	4.47%	24.82%	3.26%	0.00%	18.13%
Pacific I. Model	2.42%	20.68%	7.20%	22.08%	0.00%

D) Dementia

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	0.02%	0.63%	0.91%	0.71%
White Model	0.42%	0.00%	1.58%	1.26%	1.22%
Black Model	0.85%	0.74%	0.00%	3.00%	2.21%
Native A. Model	0.46%	0.07%	0.21%	0.00%	4.64%
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.

Figure 1 shows a very interesting result, as for two phenotypes: A) Dementia, and D) Alzheimer's disease the individual race models do not perform that differently from each other, while the other two phenotypes, B and C, there are dramatic differences in accuracy, when compared against each other. For these phenotypes, our analysis shows that identifying frailty and mild cognitive impairment might be just more complex to identify, requiring the evaluation of more phenotypes to be able to generalize in a larger context. In the additional (not shown here) leave-one-out, full, and balanced model evaluation, we also find a very interesting trend that the full and balanced models do not vary in accuracy as much for the same two phenotypes A and D, while having larger variances for B and C..

Conclusion

This 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 an input variable. This demonstrates how important it is to assess these models' performance for specific subgroups before deploying them in routine use. Certain drastic performance drops for specific groups suggest that overall well-performing algorithms could be insufficient when making classifications for certain subgroups. While most of the evaluations presented are highly automated, there is still work to be done to present proper recommendations for bias reduction in an automated and scalable way. In over 1,200 model evaluations, we have identified patterns that may indicate which phenotype algorithms are more susceptible to exhibiting bias for certain ethnic groups. Elucidating these patterns and the proper implementation of our profiling framework in APHRODITE is part of our current and future work.

References/Citations

1. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med [Internet]. 2018 Nov

1;178(11):1544–7. Available from: http://dx.doi.org/10.1001/jamainternmed.2018.3763

2. Coley RY, Johnson E, Simon GE, Cruz M, Shortreed SM. Racial/Ethnic Disparities in the Performance of Prediction Models for Death by Suicide After Mental Health Visits. JAMA Psychiatry [Internet]. 2021 Apr 28; Available from: http://dx.doi.org/10.1001/jamapsychiatry.2021.0493

3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science [Internet]. 2019 Oct 25;366(6464):447–53. Available from: http://dx.doi.org/10.1126/science.aax2342

4. Kashyap M, Seneviratne M, Banda JM, Falconer T, Ryu B, Yoo S, et al. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. J Am Med Inform Assoc [Internet]. 2020 May 6; Available from: http://dx.doi.org/10.1093/jamia/ocaa032

5. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc [Internet]. 2017 Jul 26;2017:48–57. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28815104