Characterizing subgroup performance of probabilistic phenotype algorithms within older adults: A case study for dementia, mild cognitive impairment, and Alzheimer's and Parkinson's diseases

Juan M. Banda, Ph.D[1], Nigam H. Shah, MBBS, Ph.D[2], and Vyjeyanthi S. Periyakoil, MD[3,4]

[1] Department of Computer Science, College of Arts and Sciences, Georgia State University, 25 Park Place, Suite 752, Atlanta, GA, USA
[2] Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, 1265 Welch Rd, Stanford, CA, USA
[3] Stanford Department of Medicine, 300 Pasteur Dr, Palo Alto, CA, USA
[4] VA Palo Alto health Care System, 3801 Miranda Avenue, Palo Alto, CA, USA

Corresponding author: Juan M. Banda, PhD, Department of Computer Science, College of Arts and Sciences, Georgia State University, 25 Park Place, Suite 752, Atlanta, GA, USA; jbanda@gsu.edu

MeSH terms: Phenotype, Electronic Health Records, Algorithms, Precision Medicine, Diagnosis, Computer-Assisted

Word count: 3991

ABSTRACT

Objective: Biases within probabilistic electronic phenotyping algorithms are largely unexplored. In this work, we characterize differences in sub-group performance of phenotyping algorithms for Alzheimer's Disease and Related Dementias (ADRD) in older adults.

Materials and methods: We created an experimental framework to characterize the performance of probabilistic phenotyping algorithms under different racial distributions allowing us to identify which algorithms may have differential performance, by how much, and under what conditions. We relied on rule-based phenotype definitions as reference to evaluate probabilistic phenotype algorithms created using the Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) framework.

Results: We demonstrate that some algorithms have performance variations anywhere from 3 to 30% for different populations, even when not using race as an input variable. We show that while performance differences in subgroups are not present for all phenotypes, they do affect some phenotypes and groups more disproportionately than others.

Discussion: Our analysis establishes the need for a robust evaluation framework for subgroup differences. The underlying patient populations for the algorithms showing subgroup performance differences have great variance between model features when compared to the phenotypes with little to no differences.

Conclusion: We have created a framework to identify systematic differences in the performance of probabilistic phenotyping algorithms specifically in the context of ADRD as a use case. Differences in subgroup performance of probabilistic phenotyping algorithms are not widespread nor do they occur consistently. This highlights the great need for careful ongoing monitoring to evaluate, measure, and try to mitigate such differences.

INTRODUCTION

The widespread adoption of machine learning (ML) algorithms for risk-stratification has unearthed plenty of cases of racial/ethnic biases within algorithms— from x-ray images to electronic health records (EHR) and clinical notes. [1–5]. When built without careful weightage, calibration, and bias-proofing, ML algorithms can give wrong recommendations, thereby worsening health disparities faced by communities of color. Medical researchers in fields like dermatology [6], pharmacovigilance [7], and clinical-decision support [8], to name a few, have started to examine biases inherently embedded within ML algorithms via the features used, quality of datasets, types of machine learning algorithms, and design decisions. Until the beginning of 2022, there have been over 600 published papers in PubMed, that address the evaluation and mitigation of racial bias in clinical ML models [9], with some pieces providing very insightful ideas (e.g. dividing bias in statistical and social) [10], listing challenges (e.g. adaptive learning, clinical implementation, and evaluating outcomes) [11], as well as strategies (e.g. reporting clarity, using de-noising strategies, explainability, among others) [12] on how to think about bias, where can it be present [13], and how it can be mitigated. On the implementation/deployment side, researchers have proposed how to introduce/represent these models to end-users [14] and some best practices within the field [15–18].

In the broader machine learning community, Kleinberg et al. [19] showed that a probabilistic classification to be ‘fair’ to different groups should satisfy three inherent conditions: 1) Calibration within groups, 2) Balance for the negative class, and 3) Balance for the positive class. However, these conditions cannot be satisfied all at once, which has led to the development of numerous other ‘fairness measures’ [20–22] that overlap and create confusion [23]. While most of these metrics apply to algorithms directly, they have not been analyzed in the context of medicine [24–26] until late 2019, with mixed and at times contradictory findings. When applied to medicine,other factors need to be considered, such as the clinical utility and benefit of the model [27].

Rule-based phenotyping has been the de facto method for identifying cohorts of patients belonging to any given condition/phenotype [28]. This method requires clinicians to agree on a set of clinical elements organized in logical rules that best represent the targeted phenotype. Two of the biggest disadvantages of this approach are that: 1) the rules are rigid, meaning that they do not allow patients that have missing key data points to be included, and 2) these definitions are expert-driven and very time consuming/expensive to construct. Most recently, different ML-based approaches have gained traction, because they are data driven and allow more flexibility for patient inclusion [29]. Specifically, patients are assigned probability scores, rather than a binary label. In this work, we used a probabilistic score approach to examine racial bias in EHR data of disease phenotypes that impact older adult patients.

The National Institute on Aging has defined Alzheimer’s Disease and Related Dementias (ADRD) as a series of complex brain disorders that affect millions of Americans and as having a deleterious impact on individuals, their families, long-term care facilities, health care providers, and health care systems. The negative impact of ADRD on minority older adults cannot be

overstated. In this study, we selected ADRD phenotypes as they impact all older adults, and especially in communities of color [30–33]. We utilize electronic phenotyping to characterize subgroup performance, which could lead to algorithmic biases (if any) in this context. A review of existing literature identified only one study by Straw and Wu [34]. that work Straw and Wu, present a sex-stratified analysis of machine learning models for liver disease prediction. In this work, instead of a sex-stratifed anlaysis, we build a more detailed and robust gender-stratified analysis to identify bias from a broader perspective, nicely providing an additional view of the problem than Straw and Wu. Additional prior research has examined racial bias in the context of dementia [35]. To our knowledge, this is the first study to evaluate the impact of racial subgroup performance, within probabilistic phenotyping models, on older adults in the context of mild cognitive impairment (MCI), Alzhimer's disease, and Parkinson's disease, Moreover, our study, which uses a larger EHR dataset, found very different conclusions on racial subgroup performance in dementia, thereby demonstrating the usefulness of our evaluation framework that was built for EHR data in the OMOP CDM format.

OBJECTIVE

In this study, we characterized the racial subgroup in performance of probabilistic electronic phenotyping algorithms developed from EHR datasets. Without using race as a modeling variable, we hypothesized that (1) probabilistic algorithms perform differently for different racial groups, (2) the difference in performance is tied to data availability for different racial groups, and(3) not all algorithms show the same level of racial subgroup performance differences.

MATERIALS AND METHODS

Dataset

The dataset utilized for this work is from de-identified data from the Stanford Medicine Research data Repository (STARR), consisting of over three million patients with clinical data from 2008 until the end of 2018. This dataset has been converted to the OMOP CDM version 5.3 and used in multiple OHDSI studies over the years. Table 1 shows the overall demographics of the dataset. Of particular interest is the underlying racial distribution, which will be highlighted throughout the rest of this study.

Table 1: Demographics of patients from 2008-2018

	Overall		
		n= 3,113,080	
Gender (%)			
	Missing	5,806	(0.2)
	Female	1,673,410	(53.8)
	Male	1,433,864	(46.1)
Race (%)			

	Asian	303,800	(9.8)
	Black	97,399	(3.1)
	Native American	6,819	(0.2)
	Other	393,466	(12.6)
	Pacific Islander	23,142	(0.7)
	Unknown	1,071,563	(34.4)
	White	1,216,891	(39.1)
Ethnicity (%)			
	Hispanic/Latino	338,820	(10.9)
	Non-Hispanic	1,602,753	(51.5)
	Unknown	1,171,507	(37.6)
	Age (mean (sd))	46.7	(25.11)

Rule-based algorithms as our gold-standard

We used the rule-based phenotype algorithms validated and available on the HDR UK Phenotype library [36] for the following conditions: dementia [37] and Parkinson's disease [38]. We adapted them into OHDSI ATLAS cohort definitions for their application on our OMOP-converted data. The rule-based phenotype definitions for mild cognitive impairment were adapted from Jongsiriyanyong and Limpawattana [39], and for Alzheimer's disease we used the clinical definition by Holmes [40].

Manually curated and clinically validated sets of patients are more robust but also significantly more time and resource intensive. Thus community-approved rule-based definitions from organizations like PheKB [41] and CALIBER (now HDR UK Phenotype library) [42] have become the next best thing in order to computationally identify phenotypes. These rule-based definitions, while clinically validated, are quite rigid and show less flexibility than other approaches [28], this leads to many potential patients being excluded if certain codes or conditions are just not presented in their health record because of lack of coding or errors.

Probabilistic phenotypes with APHRODITE

Instead of relying on rigid rules to define medical condition phenotypes, newer data driven approaches leverage machine learning to build statistical models to classify patients. These approaches have gained traction in the last few years [28], because they allow subjects to have a degree of probability of belonging to a phenotype, making them more flexible and able to catch people that have clinical codes missing or incomplete data. The methodology used for this work, APHRODITE, was designed with this flexibility in mind. Specifically, it relies on building statistical models for phenotypes based on an initial cohort of patients selected using high-precision keywords/clinical codes. The models were built using weak supervision where

the patient's entire clinical record up until the first appearance of the selected keyword/clinical code and. use weak supervision, where only an initial keyword/clinical code is needed and everything else is data driven. This approach was introduced by Agarwal et al. [43] and was made into an R package [44] which works on standardized data to the OMOP CDM format.

Experimental framework

In order to identify racial subgroup performance variations within the probabilistic models, we built the following steps into a framework that we can later re-use for a wider variety of phenotypes. We started by selecting three of the most popular classical machine learning models to evaluate: LASSO [45], since regression-based classifiers are widely used for statistical learning purposes with EHR/medical data [46], Random Forest (RF) [47], and support vector machines (SVM) [48]. Note that the three models listed above are the ones supported by default by APHRODITE; however,any model supported by caret R package can also be included [49]. Next, we selected matched cases and controls to build our probabilistic models. The patients were matched by age, race, gender, and length of clinical record. We then stratified by the patient's race for our multi-pronged evaluation, which consists of: traditional model (all races merged together), balanced model (we balance based on equal distribution of patients for each race), single race only model, and the leave-one-out combinations, which take one race out of the model building process in a systematic way. Following our usual practice, we used 75% of the data to train the model and a 25% unseen set to test the model, in addition to five-fold cross validation.

For evaluation, we used the traditional metrics: accuracy, which is the fraction of assignments the model identified correctly; sensitivity, which is the proportion of positives that are correctly identified; and specificity, which measures the proportion of negatives that are correctly identified. In addition, we used variation in order to measure difference between models in the following way:

$$\text{Variation} = \left| 1 - \frac{\text{current model}}{\text{base comparison model}} \right|$$

This measurement allowed us to evaluate the first three metrics in a similar context and show how different models are compared with each other. Note that while phenotyping algorithm performance is important, this is not the key point of this work, we present general model classification accuracy in order to put performance variance between phenotypes in context.

RESULTS

Phenotyping algorithms

First, we checked that the probabilistic phenotyping algorithms performed well when compared against their rule-based definitions. **Table 2** shows APHRODITE's performance to select and identify the 'gold-standard' patients identified by the rule-based phenotype definitions.

Table 2: Rule-based and APHRODITE phenotype selection overlap.

Phenotype	Cases identified by rule-based definition	Cases identified by APHRODITE keywords	Initial Overlap (%)	Classified by APHRODITE model (prob over 90%)	Total Overlap (%)
Dementia	13,213	16,998	95.95%	471	99.52%
Mild Cognitive Impairment	7,915	8,292	94.14%	399	99.18%
Alzheimer's disease	11,401	12,828	89.04%	1,137	99.02%
Parkinson's disease	5,989	6,644	79.50%	896	94.46%

Our results show that APHRODITE and its probabilistic models are successful at identifying almost the same patients for each phenotype when compared to the rule-based definition, which served as our gold-standard.

Table 3 shows the demographics of the patients identified for each phenotype. These results had a considerable impact on the resulting models and downstream evaluations.

Table 3: Patient demographics for the evaluated phenotypes.

Phenotype	Cases	Controls	Gender		Race					
			Female	Male	Asian	Black	Native Amer.	Pacific Island.	White	Unknown
Dementia	16,998	16,998	56.22%	43.78%	11.07%	4.96%	0.27%	0.80%	60.33%	22.57%
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%
Parkinson's disease	6,644	6,644	39.30%	60.70%	12.06%	1.34%	0.23%	0.32%	63.11%	22.95%

A few things to note are that: 1) the racial distributions for some of the categories like Native American and Pacific Islanders are very low; 2) there were many patients listed as unknown race, which were removed from our evaluation.

Our evaluation framework produced over 1,200 plots and charts evaluating the performance of probabilistic models built under multiple conditions. **Figure 1** presents model classification accuracies for the four phenotypes and the three machine learning algorithms we utilized (i.e., 15 parameters) for the twelve different data subsets evaluated. This figure sets the precedent of the importance of the model variance evaluation and how it will change models from being potentially useful, to highly unreliable.

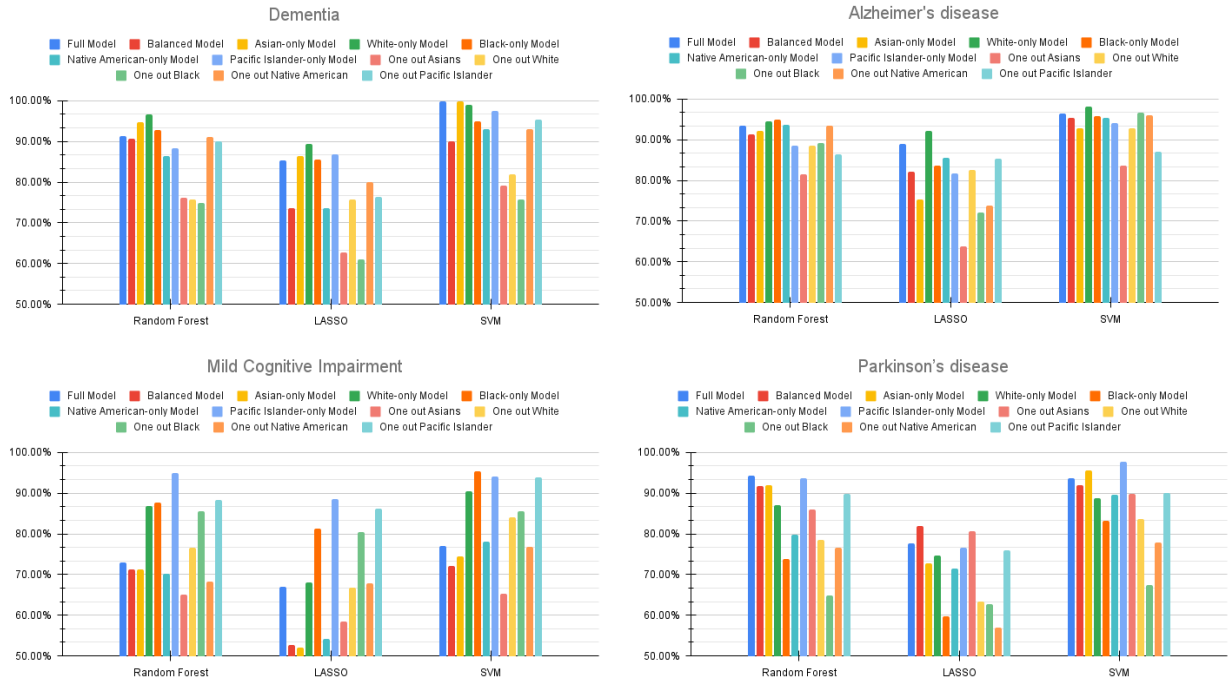

Figure 1. Model Classification Accuracy Result for all machine learning algorithms and all phenotypes.

These results demonstrate that most full models (i.e., the classical ones), which are the purple bars, have 70% to 90% classification accuracy for the given phenotype. These results are only for illustration purposes and to put the following figures of model variance in perspective. We are not trying to find the best performing models in general, but rather show their bias, when races are stratified.

Evaluation scenario one: Building models with individual racial subgroups

In Figures 2 and 3 we show the classification and sensitivity variance for the random forest models between our models built for individual races and compared across all races evaluation. For example we built models using only White patients, and compared their performance when classifying patients from all other races. We used random forest as our choice algorithm to illustrate our results due to its solid average performance during our experimentation, and due to the more explainable nature of its models. Note that the same plots for specificity, as well as for the LASSO and SVM classifiers are provided as part of the supplemental appendix.

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%

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%

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%

Parkinson's disease

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	22.34%	15.32%	10.62%	1.05%
White Model	6.15%	0.00%	5.12%	7.51%	18.48%
Black Model	3.35%	26.38%	0.00%	20.59%	21.97%
Native A. Model	12.91%	3.32%	15.65%	0.00%	0.56%
Pacific I. Model	11.16%	6.84%	5.11%	20.17%	0.00%

Figure 2. Classification accuracy variance for the Random Forest algorithm for our four phenotypes. In this heatmap, the higher tone of red a cell has, represents bigger variance from the baseline, or in other words more bias.

Figure 2 shows some very interesting results for the Dementia and Alzheimer's disease phenotypes, illustrating that the variance between models is not that pronounced (less than 8% at the worst case). This shows that models built with the individual races are generalizable enough across races, at least for these two phenotypes. For the remaining three phenotypes the variance increases up to 27%, rendering a classification model with an accuracy of less than 80%, almost equal to random picking. These results showcase the need to evaluate these combinations carefully, particularly before deployment of any phenotyping model.

Dementia

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	5.85%	6.52%	1.85%	7.94%
White Model	0.29%	0.00%	2.68%	0.58%	0.53%
Black Model	6.58%	7.14%	0.00%	1.53%	12.95%
Native A. Model	16.11%	16.39%	16.96%	0.00%	26.47%
Pacific I. Model	6.65%	6.33%	5.77%	9.67%	0.00%

Alzheimer's disease

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	5.17%	5.99%	1.41%	4.29%
White Model	0.89%	0.00%	1.74%	5.37%	0.55%
Black Model	4.34%	4.45%	0.00%	4.40%	7.19%
Native A. Model	5.94%	5.34%	4.73%	0.00%	1.44%
Pacific I. Model	1.97%	1.43%	0.63%	3.12%	0.00%

Mild Cognitive Impairment

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	22.69%	4.32%	7.20%	7.18%
White Model	10.40%	0.00%	1.59%	4.84%	7.97%
Black Model	4.29%	11.15%	0.00%	15.70%	3.52%
Native A. Model	6.91%	23.97%	1.74%	0.00%	17.34%
Pacific I. Model	0.65%	6.74%	9.90%	14.46%	0.00%

Parkinson's disease

	Asian	White	Black	Native American	Pacific Islander
Asian Model	0.00%	14.09%	11.47%	1.57%	5.66%
White Model	5.84%	0.00%	18.08%	0.47%	12.12%
Black Model	3.15%	18.30%	0.00%	10.37%	21.44%
Native A. Model	17.03%	9.48%	15.23%	0.00%	18.63%
Pacific I. Model	19.05%	5.62%	9.76%	8.47%	0.00%

Figure 3. Classification sensitivity variance for the random forest algorithm for our four phenotypes. In this heatmap, the higher tone of red a cell has, represents bigger variance from the baseline, or in other words more bias.

The sensitivity variance for the models on Figure 3 shows a similar trend as Figure 2–i.e., less variance for the Dementia and Alzheimer's phenotypes. However, there is a considerable increase in the variance of the Native American model, most likely due to the fact that this racial

group is highly underrepresented in the dataset used; thus this model is unable to properly generalize across races, particularly when measuring sensitivity.

Evaluation scenario two: Full models, balanced models, and leave one-out

We now switch to evaluate the models in a more traditional sense of using a model with all data available–one that balances the classes but limits the number of samples to the minimum available in any given class. Note that we did not use SMOTE [50] or any sampling techniques in this work, as this is not ideal when using clinical data [51,52], since depending on the method used, it adds non-representative extra data. The other scenarios we evaluated include leaving one class out in the model building process. We then applied the built models to the individual classes of patients in the testing set (fully unseen patients). Figures 4 and 5 report these results for classification accuracy and sensitivity. Additional figures for the other machine learning learning models and metrics are available in the supplemental appendix.

Dementia

	Asian	White	Black	Native American	Pacific Islander
Full Model	3.85%	2.99%	2.15%	3.21%	4.78%
Balanced Model	2.70%	1.97%	2.21%	2.40%	1.54%
One out Asians	14.50%	21.04%	2.83%	20.48%	23.42%
One out White	20.59%	25.35%	20.46%	4.52%	22.24%
One out Black	6.81%	19.06%	27.33%	4.26%	5.70%
One out Native A.	12.50%	14.52%	13.58%	1.32%	1.05%
One out Pacific I.	19.52%	0.51%	16.52%	12.53%	5.48%

Alzheimer's disease

	Asian	White	Black	Native American	Pacific Islander
Full Model	1.81%	2.11%	1.61%	1.19%	2.17%
Balanced Model	1.49%	0.46%	1.12%	3.57%	2.56%
One out Asians	19.06%	24.40%	16.23%	5.05%	3.10%
One out White	16.97%	3.64%	16.84%	13.94%	18.14%
One out Black	0.07%	6.09%	2.75%	24.00%	1.74%
One out Native A.	13.23%	13.61%	2.61%	9.21%	4.16%
One out Pacific I.	12.77%	2.14%	7.36%	21.07%	1.57%

Mild Cognitive Impairment

	Asian	White	Black	Native American	Pacific Islander
Full Model	11.71%	10.14%	10.20%	12.28%	11.51%
Balanced Model	11.93%	10.97%	11.33%	13.21%	9.38%
One out Asians	6.74%	18.22%	23.02%	5.68%	5.32%
One out White	13.72%	10.09%	0.78%	20.31%	5.57%
One out Black	16.78%	12.22%	19.38%	5.05%	7.28%
One out Native A.	19.99%	21.07%	4.02%	1.10%	3.88%
One out Pacific I.	18.53%	1.83%	4.29%	9.24%	3.18%

Parkinson's disease

	Asian	White	Black	Native American	Pacific Islander
Full Model	11.56%	12.47%	12.89%	11.14%	11.64%
Balanced Model	11.52%	11.82%	11.44%	10.99%	10.86%
One out Asians	23.87%	14.08%	7.32%	18.74%	11.17%
One out White	1.37%	0.69%	0.47%	12.45%	8.22%
One out Black	14.27%	6.31%	7.75%	7.00%	21.88%
One out Native A.	19.67%	14.89%	0.68%	3.63%	22.28%
One out Pacific I.	20.60%	11.56%	2.62%	2.39%	3.87%

Figure 4. Classification accuracy variance for the Random Forest algorithm for our four phenotypes. In this heatmap, the higher tone of red a cell has, represents bigger variance from the baseline, or in other words more bias.

We again see very similar patterns for the Dementia and Alzheimer's disease phenotypes, where the variance is quite low. One thing to note is that there is always variance here as the calculation is performed on the baseline model performance only classifying the given class (on the unseen test set) versus the baseline model performance of all classes on the unseen test set. One very interesting result here is that for the phenotype models with less subgroup variance, taking out certain classes brings their overall performance down by considerable amounts (up to 26%) sometimes, which could be mostly due to removing the data-dominant race. In other scenarios the accuracy variance is not that high, particularly when classifying the races with very small representation in the original models.

Dementia

	Asian	White	Black	Native American	Pacific Islander
Full Model	0.07%	1.46%	1.91%	2.99%	0.99%
Balanced Model	1.38%	1.83%	2.50%	3.24%	3.19%
One out Asians	14.97%	8.21%	4.56%	2.18%	9.96%
One out White	17.39%	6.44%	5.28%	7.64%	16.39%
One out Black	11.47%	14.05%	16.99%	4.37%	0.87%
One out Native A.	8.15%	12.72%	14.24%	19.47%	12.90%
One out Pacific I.	6.49%	4.32%	16.16%	8.86%	5.76%

Alzheimer's disease

	Asian	White	Black	Native American	Pacific Islander
Full Model	2.87%	3.71%	0.29%	2.81%	1.62%
Balanced Model	1.43%	1.84%	0.78%	1.83%	1.77%
One out Asians	5.17%	25.07%	0.51%	11.03%	21.51%
One out White	4.92%	25.38%	13.51%	26.07%	0.75%
One out Black	7.29%	5.72%	7.06%	0.20%	13.08%
One out Native A.	20.41%	0.64%	1.26%	3.51%	6.94%
One out Pacific I.	4.81%	18.63%	8.36%	14.59%	16.53%

Mild Cognitive Impairment

	Asian	White	Black	Native American	Pacific Islander
Full Model	10.44%	11.87%	11.25%	3.04%	10.57%
Balanced Model	10.67%	11.09%	10.80%	11.65%	11.09%
One out Asians	3.05%	6.29%	9.53%	5.11%	6.95%
One out White	16.72%	17.71%	3.85%	17.91%	15.87%
One out Black	5.34%	22.89%	8.83%	23.72%	3.51%
One out Native A.	9.64%	6.09%	5.57%	7.42%	2.70%
One out Pacific I.	8.01%	14.12%	11.29%	0.54%	4.97%

Parkinson's disease

	Asian	White	Black	Native American	Pacific Islander
Full Model	11.83%	10.77%	11.07%	10.39%	13.38%
Balanced Model	8.60%	13.33%	13.52%	14.37%	11.79%
One out Asians	20.44%	19.76%	12.39%	10.02%	6.91%
One out White	7.59%	28.32%	15.24%	3.32%	6.00%
One out Black	13.65%	9.87%	0.77%	1.11%	8.87%
One out Native A.	7.57%	7.05%	23.57%	10.32%	3.18%
One out Pacific I.	9.27%	19.20%	12.92%	14.67%	11.03%

Figure 5. Classification sensitivity variance for the Random Forest algorithm for our four phenotypes. In this heatmap, the higher tone of red a cell has, represents bigger variance from the baseline, or in other words more bias.

Figure 5 shows that classification sensitivity variance is also affected in a similar way as for the classification accuracy in almost the exact same way with the variance trends being very similar. This also reinforces the original finding, namely that two of the phenotypes show small variance, across our experimental evaluation and the other three have different, but increasing, degrees of variance. These results demonstrate that probabilistic phenotype models need to be carefully examined and improved before they can be used in a clinical setting.

DISCUSSION

We designed our experimental framework to provide a fully-automated and standardized (on top of the OMOP CDM and using APHRODITE R package) way to demonstrate if any given probabilistic phenotyping model has racial subgroup variance and estimate how much. As shown in the results sections, we have two different evaluation scenarios, which build probabilistic models in different stratified ways to provide flexibility and insight into how the differently built models will vary given popular machine learning metrics. Note that any identified anomaly in subgroup performance does not automatically translate into the algorithm leading 'harming' subgroups, as addressing or 'fixing' those anomalies might actually produce worse performing algorithms for all subgroups as found by Pfohl et al. in [53]. In some cases having different algorithms for subgroups or a human in the loop might [54] be a better approach to not affect any group of patients.

For the phenotypes of our case study, the first scenario clearly demonstrates that two phenotypes: dementia and Alzheimer's disease present the least amount of sub group variance, (Figures 2 and 3), both in terms of classification accuracy and sensitivity. This is highlighted by their two variance figures showing the least amount of red cells. These figures are designed as heatmaps to visually highlight the severity of the variance and place the attention of the

researcher on the more relevant sections. Racial representation in those cohorts could be one possible reason these phenotypes show less subgroup variance, but while they have some of the highest numbers of cases used for training (Table 2), their racial representation is nearly the same as in the whole dataset. An interesting observation is that accuracy shows very little variance whereas the sensitivity results show a higher amount of it. This might indicate that while the stratified models do well as a whole, there might be additional small variance when trying to detect the positive class. However, this can also be explained as an artifact of building the model with the Native American patients, which had the least representation in the full dataset (> 0.25%). A common solution to address underrepresentation has been to oversample this class, or undersample some of the others. However, our second scenario shows that this does not significantly improve the level of subgroup variance, as other researchers have shown, at least for predictive tasks [51,52]. Rather, we recommend using federated learning, as this is starting to be more accepted in large research networks [55], or use ensembles using multiple datasets, when available within a single site or research facility [56]. The second evaluation scenario shows that the full model and the balanced model perform very consistently for the phenotypes of dementia and Alzheimer's disease, which have less subgroup variance. However, it also shows very striking results on the leave-one-out models, particularly when removing the racial groups with the larger representations, and when removing a particular racial group and then trying to classify only patients of that group (Figures 4 and 5). These results show the need to consider such detailed analyses before trying to use any of these models in clinical practice.

Regarding the phenotypes—mild cognitive impairment, and Parkinson's disease— with more subgroup variance, we observed some very dramatic variance changes (average of 10%) between most of the experimental models. This indicates that those phenotypes are quite sensitive to any of the racial groups being removed, particularly shown by the leave-one-out models from the second scenario (Figures 4 and 5). These figures also show that even for the full-model and balanced scenarios, predicting on individual classes brings considerable accuracy and sensitivity variance. These findings translate to how sensitive these models are to any type of shift in the underlying dataset that is used to train the model and how to evaluate them. These results strongly demonstrate the need for rigorous experimental evaluation before any kind of deployment or testing in production environments (e.g. hospitals). While there are plenty of experimental evaluations to analyze, our framework automates the work for researchers, and it only needs human interpretation of its findings.

The limitations of this work are the following: we evaluated three machine learning algorithms based on their popularity and level of use within the field. However, with new algorithms constantly being introduced, the results could vary dramatically when other algorithms are introduced. We decided to keep the same algorithms as in our previous works [44,57] since we know how those perform to build probabilistic models using APHRODITE. The flexibility behind APHRODITE, allows for other models to be evaluated within the presented framework as long as they have an R package available. We decided to only evaluate the variance metric as it gives a stronger and more interpretable signal on how the model differs from each other. For this study we needed to keep the number of evaluations and experiments to a reasonable

amount to be able to explain this work and its merits. However, any other metric can be configured into our framework. Lastly, our case study phenotypes were evaluated on a single dataset, and, in future work, we plan to fully leverage the OHDSI community to conduct a network study examining racial bias of phenotypic algorithms [58,59]. One major item to note is that self-reported race is usually error prone and very often incomplete (missing in up to 23% of the patients selected for the phenotypes evaluated), these factors could lead to some of the results being artifacts of this phenomenon.

CONCLUSION

As we have demonstrated in this work, subgroup performance variance can certainly be found in probabilistic phenotype algorithms that categorize older adults, particularly for phenotypes like mild cognitive impairment, and Parkinson's disease. As a result of this subgroup variance, models perform up to 30% worse under our multiple model building scenarios. Thus it is critical for institutions to extensively test and rigorously evaluate their phenotyping models. We found that some phenotypes like dementia and Alzheimer's disease were more resistant to this subgroup variance as indicated by their very small variance under all of our testing scenarios, meaning that these models could be potentially used safely. Rigorous testing allows researchers to be more confident of the performance of these models under different racial distributions. Our main contribution is the framework to fully automate this process when the institution has data in the OMOP CDM and can run our extension to the APHRODITE package. This work is particularly important as biomedical scientists and medical professionals strive to make informed conclusions and diagnoses of older adult patients.

FUNDING

This work was supported by the National Institute on Aging of the National Institutes of Health (3P30 AG059307-02S1)

AUTHOR CONTRIBUTIONS

Conception and design: All authors. Data analysis: JMB. Data interpretation: JMB. Drafting the manuscript: All authors. Revising the manuscript: All authors. Approval of submitted version: All authors. Accountability of own contributions: All authors.

CONFLICTS OF INTEREST STATEMENTS

The authors declare that they have no competing interests.

DATA AVAILABILITY

The EHR data used in this study cannot be shared for ethical/privacy reasons. All code is available in the following location: https://github.com/OHDSI/Aphrodite

ACKNOWLEDGEMENTS

The authors would like to thank Jessica Moon, PhD, PMP for her comments and proofreading of this manuscript.

REFERENCES

1 Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA J Ethics* 2019;**21**:E167–79. doi:10.1001/amajethics.2019.167

2 Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**:447–53. doi:10.1126/science.aax2342

3 Seyyed-Kalantari L, Liu G, McDermott M, *et al.* CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: *Biocomputing 2021*. WORLD SCIENTIFIC 2020. 232–43. doi:10.1142/9789811232701_0022

4 Burlina P, Joshi N, Paul W, *et al.* Addressing Artificial Intelligence Bias in Retinal Diagnostics. *Transl Vis Sci Technol* 2021;**10**:13. doi:10.1167/tvst.10.2.13

5 Thompson HM, Sharma B, Bhalla S, *et al.* Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *J Am Med Inform Assoc* 2021;**28**:2393–403. doi:10.1093/jamia/ocab148

6 Daneshjou R, Smith MP, Sun MD, *et al.* Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatol* 2021;**157**:1362–9. doi:10.1001/jamadermatol.2021.3129

7 Kompa B, Hakim JB, Palepu A, *et al.* Artificial Intelligence Based on Machine Learning in Pharmacovigilance: A Scoping Review. *Drug Saf* 2022;**45**:477–91. doi:10.1007/s40264-022-01176-1

8 Čartolovni A, Tomičić A, Lazić Mosler E. Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. *Int J Med Inform* 2022;**161**:104738. doi:10.1016/j.ijmedinf.2022.104738

9 Huang J, Galal G, Etemadi M, *et al.* Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. *JMIR Med Inform* 2022;**10**:e36388. doi:10.2196/36388

10 Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA* Published Online First: 22 November 2019. doi:10.1001/jama.2019.18058

11 DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc* 2020;**27**:2020–3. doi:10.1093/jamia/ocaa094

12 Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Commun Med* 2021;**1**:25. doi:10.1038/s43856-021-00028-w

13 Gianfrancesco MA, Tamang S, Yazdany J, *et al.* Potential Biases in Machine Learning

Algorithms Using Electronic Health Record Data. *JAMA Intern Med* 2018;**178**:1544–7. doi:10.1001/jamainternmed.2018.3763

14 Sendak MP, Gao M, Brajer N, *et al.* Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med* 2020;**3**:41. doi:10.1038/s41746-020-0253-3

15 de Hond AAH, Leeuwenberg AM, Hooft L, *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med* 2022;**5**:2. doi:10.1038/s41746-021-00549-7

16 Rajpurkar P, Chen E, Banerjee O, *et al.* AI in health and medicine. *Nat Med* 2022;**28**:31–8. doi:10.1038/s41591-021-01614-0

17 Luo W, Phung D, Tran T, *et al.* Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* 2016;**18**:e323. doi:10.2196/jmir.5870

18 Liu X, Glocker B, McCradden MM, *et al.* The medical algorithmic audit. *Lancet Digit Health* 2022;**4**:e384–97. doi:10.1016/S2589-7500(22)00003-6

19 Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany 2017. doi:10.4230/LIPIcs.ITCS.2017.43

20 Chouldechova A, Roth A. The Frontiers of Fairness in Machine Learning. arXiv [cs.LG]. 2018.http://arxiv.org/abs/1810.08810

21 Beutel A, Chen J, Doshi T, *et al.* Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: : Association for Computing Machinery 2019. 453–9. doi:10.1145/3306618.3314234

22 Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. *Commun ACM* 2020;**63**:82–9. doi:10.1145/3376898

23 Castelnovo A, Crupi R, Greco G, *et al.* A Clarification of the Nuances in the Fairness Metrics Landscape. arXiv [cs.LG]. 2021.http://arxiv.org/abs/2106.00467

24 Xu J, Xiao Y, Wang WH, *et al.* Algorithmic fairness in computational medicine. bioRxiv. 2022. doi:10.1101/2022.01.16.21267299

25 McCradden MD, Joshi S, Mazwi M, *et al.* Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health* 2020;**2**:e221–3. doi:10.1016/S2589-7500(20)30065-0

26 Chen RJ, Chen TY, Lipkova J, *et al.* Algorithm Fairness in AI for Medicine and Healthcare. arXiv [cs.CV]. 2021.http://arxiv.org/abs/2110.00603

27 Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med* 2018;**378**:981–3. doi:10.1056/NEJMp1714229

28 Banda JM, Seneviratne M, Hernandez-Boussard T, *et al.* Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* 2018;**1**:53–68. doi:10.1146/annurev-biodatasci-080917-013315

29 Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *J Am Med Inform Assoc* Published Online First: 12 October 2017. doi:10.1093/jamia/ocx110

30 Fredriksen-Goldsen KI, Kim H-J, Barkan SE, *et al.* Health disparities among lesbian, gay, and bisexual older adults: results from a population-based study. *Am J Public Health* 2013;**103**:1802–9. doi:10.2105/AJPH.2012.301110

31 Dunlop DD, Manheim LM, Song J, *et al.* Gender and ethnic/racial disparities in health care utilization among older adults. *J Gerontol B Psychol Sci Soc Sci* 2002;**57**:S221–33. doi:10.1093/geronb/57.4.s221

32 Ward JB, Gartner DR, Keyes KM, *et al.* How do we assess a racial disparity in health? Distribution, interaction, and interpretation in epidemiological studies. *Ann Epidemiol* 2019;**29**:1–7. doi:10.1016/j.annepidem.2018.09.007

33 Johnson KS. Racial and ethnic disparities in palliative care. *J Palliat Med* 2013;**16**:1329–34. doi:10.1089/jpm.2013.9468

34 Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health Care Inform* 2022;**29**. doi:10.1136/bmjhci-2021-100457

35 Gianattasio KZ, Ciarleglio A, Power MC. Development of Algorithmic Dementia Ascertainment for Racial/Ethnic Disparities Research in the US Health and Retirement Study. *Epidemiology* 2020;**31**:126–33. doi:10.1097/EDE.0000000000001101

36 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health* 2019;**1**:e63–77. doi:10.1016/S2589-7500(19)30012-3

37 Phenotype Library. https://phenotypes.healthdatagateway.org/phenotypes/PH148/version/296/detail/ (accessed 15 Jun 2022).

38 Phenotype library. https://phenotypes.healthdatagateway.org/phenotypes/PH77/version/154/detail/ (accessed 15 Jun 2022).

39 Jongsiriyanyong S, Limpawattana P. Mild Cognitive Impairment in Clinical Practice: A Review Article. *Am J Alzheimers Dis Other Demen* 2018;**33**:500–7. doi:10.1177/1533317518791401

40 Holmes C. Genotype and phenotype in Alzheimer's disease. *Br J Psychiatry* 2002;**180**:131–4. doi:10.1192/bjp.180.2.131

41 Kirby JC, Speltz P, Rasmussen LV, *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016;**23**:1046–52. doi:10.1093/jamia/ocv202

42 Denaxas S, Gonzalez-Izquierdo A, Direk K, *et al.* UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Inform Assoc* 2019;**26**:1545–59. doi:10.1093/jamia/ocz105

43 Agarwal V, Podchiyska T, Banda JM, *et al.* Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc* 2016;**23**:1166–73. doi:10.1093/jamia/ocw028

44 Banda JM, Halpern Y, Sontag D, *et al.* Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc* 2017;**2017**:48–57.https://www.ncbi.nlm.nih.gov/pubmed/28815104

45 Tibshirani R. Regression Shrinkage and Selection Via the Lasso. In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*. 1994. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574 (accessed 13 Jun 2022).

46 Jones AM. Models for Health Care. In: Michael P. Clements and David F. Hendry, ed. *The Oxford Handbook of Economic Forecasting*. Oxford University Press 2011. doi:10.1093/oxfordhb/9780195398649.013.0024

47 Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32. doi:10.1023/A:1010933404324

48 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. doi:10.1007/BF00994018

49 Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw* 2008;**28**:1–26. doi:10.18637/jss.v028.i05

50 Chawla NV, Bowyer KW, Hall LO, *et al.* SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002;**16**:321–57. doi:10.1613/jair.953

51 van den Goorbergh R, van Smeden M, Timmerman D, *et al.* The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc* Published Online First: 10 June 2022. doi:10.1093/jamia/ocac093

52 Pfohl SR, Zhang H, Xu Y, *et al.* A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Sci Rep* 2022;**12**:3254. doi:10.1038/s41598-022-07167-7

53 Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform* 2021;**113**:103621. doi:10.1016/j.jbi.2020.103621

54 Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA* 2018;**319**:19–20. doi:10.1001/jama.2017.19198

55 Xu J, Glicksberg BS, Su C, *et al.* Federated Learning for Healthcare Informatics. *Int J Healthc Inf Syst Inform* 2021;**5**:1–19. doi:10.1007/s41666-020-00082-4

56 Reps JM, Williams RD, Schuemie MJ, *et al.* Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. *BMC Med Inform Decis Mak* 2022;**22**:142. doi:10.1186/s12911-022-01879-6

57 Kashyap M, Seneviratne M, Banda JM, *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* Published Online First: 6 May 2020. doi:10.1093/jamia/ocaa032

58 Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016;**113**:7329–36. doi:10.1073/pnas.1510502113

59 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;**216**:574–8.https://www.ncbi.nlm.nih.gov/pubmed/26262116