Supplemental Online Content

Yuan C, Linn KA, Hubbard RA. Algorithmic fairness of machine learning models for Alzheimer disease progression. *JAMA Netw Open*. 2023;6(11):e2342203. doi:10.1001/jamanetworkopen.2023.42203

This supplemental material has been provided by the authors to give readers additional information about their work.

eMethods

***Training details*:** All experiments in this study were conducted on One Nvidia RTX 3090 GPU, one Inter i9-12900F CPU with 16 cores and 32G RAM. Pytorch 1.0 and Python 2.7 were used to define all models and training procedures.

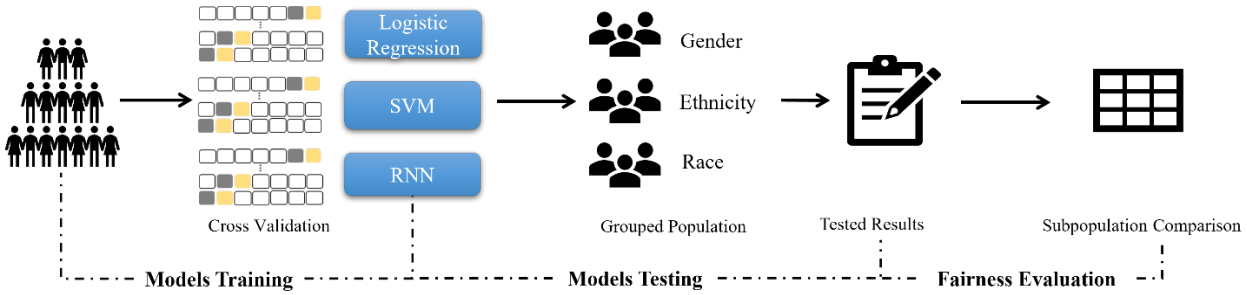
For logistic regression, we used the implementation of logistic regression in sklearn library[48] for a three-class classification problem. We used L2 penalty (ridge regression) of weights, 1000 iterations, and the default solver lbfgs[49] to learn the weights. To train a SVM model, we followed the work from Nguyen.[35] Since SVM accepts fixed length feature vectors and it cannot handle subjects with different number of inputs timepoints. We trained different SVM models using 1 to 4 input timepoints (spaced 6 months apart) to predict the future observations. We trained 40 SVM models on four input timepoints (1, 2, 3 or 4 input timepoints) to predict clinical diagnosis as outcome for 10 future predictions (6, 12, 18, …, 60 months), in which 4×10 = 40 SVM models. The four timepoints were validated as the best settings in Nguyen's work.[35] The maximum iteration for training SVM is set to 10^5. The SVM model utilized the radial basis function kernel, and the process of tuning hyperparameters remained consistent with the approach described in Nguyen's paper [35].

We adapted the minimalRNN [3] for predicting disease progression. The input x_t to each RNN cell comprised the diagnosis s_t and continuous variables g_t, $x_t = [s_t, g_t]$. The hidden state h_t was a combination of the previous hidden state h_{t-1} and the transformed input u_t, $u_t = tanh\ tanh\ (W_x x_t)$, $h_t = f_t \odot h_{t-1} + (1 - f_t) \odot u_t$. The forget gate f_t weighed the contribu tions of the previous hidden state h_{t-1} and current transformed input u_t toward the current hidden state h_t , $f_t = \sigma(U_h h_{t-1} + W_u u_t)$. The model predicted the next month diagnosis $\hat{s}_{t+1}$ and continuous variables using the hidden state h_t, $\hat{s}_{t+1} = (W_s h_t)$, $\hat{g}_{t+1} = W_g h_t + g_t$. $\odot$ and σ denote element-wise product and the sigmoid function respectively. To train the RNN model, we set batch size as 128 and epoch number aS200. We use Adam[50] as the optimizer with learning rate of 5×10^{-4}, the value of β1 as 0.9 and β2 as 0.999, and weight decay as 5×10^{-7} to avoid overfitting. As in Nguyen's paper [35] , we used an unweighted sum of cross-entropy loss for categorical variable (diagnosis stage) and MAE loss for the continuous variables.

We used cross-validation for model selection and evaluation. The stratification was used during the partitioning to return stratified folds with non-overlapping groups and the folds are made by preserving the percentage of samples for each class. The selected model was determined by the best accuracy on validation data set. As the focus of this paper is on assessing the fairness in machine learning models on predicting AD as opposed to risk prediction model development, we do not report details of the predictors and model performance during the training phase in detail. Further description of variables and model performance can be found in Nguyen's work.[35]

Population: We employed the dataset provided by the TADPOLE challenge [31] where the organisers provided participants with a standard ADNI-derived dataset to train algorithms, removing the need for participants to pre-process the ADNI data or merge different spreadsheets.

***Model performance*:** Following the same evaluation of model performance in Nguyen's work,[35] diagnosis classification accuracy was evaluated using the multiclass area under the operating curve (mAUC)51 and balanced class accuracy (BCA) metrics. The mAUC was computed as the average of three two-class AUC (AD vs not AD, MCI vs. not MCI, and CN vs not CN). mAUC is independent of the group sizes and gives an overall measure of classification ability that accounts for relative likelihoods assigned to each class. The BCA for each class was computed as $\frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$ and the overall BCA was given by the mean of all the balanced accuracies for every cognitive trajectory. BCA considers the accuracy of the most likely classification. For both mAUC and BCA metrics, higher values indicate better performance. The performance was evaluated by averaging the results across 10 test sets for logistic regression, SVM, and RNN. The results for the three models are shown in Table S2.



eFigure 1. An overview of the model pipeline. Model Training: we trained three ML models using cross-validation from entire populations to predict the progression to AD; Model Testing: we tested three models across the different grouped populations, including gender, ethnicity, and race; Fairness Evaluation: we assessed the fairness metrics on test results.

eTable 1. Summary statistics for protected attributes and predictor variables stratified by cognitive functioning trajectory (CN-AD, MCI-CN, AD-stable and AD-MCI) for trajectories excluded from fairness analysis due to small sample size.

	CN-AD	MCI-CN	AD-stable	AD-MCI
Outcome Number	24	143	337	3
Protected Attributes (N (%))				
Gender				
Female	14 (58%)	82 (57%)	151 (45%)	1 (33%)
Male	10 (42%)	61 (43%)	186 (55%)	2 (67%)
Ethnicity				
Hispanic	0 (0%)	7 (5%)	14 (4%)	0 (0%)
Non-Hispanic	24 (100%)	136 (95%)	323 (96%)	3 (100%)
Race				
Asian	0 (0%)	0 (0%)	7 (2%)	0 (0%)
Black	1 (4%)	4 (3%)	14 (4%)	0 (0%)
White	23 (96%)	139 (97%)	312 (93%)	3 (100%)
Others	0 (0%)	0 (0%)	4 (1%)	0 (0%)
Predictors (mean (SD))				
Clinical Dementia Rating Scale	1.99 (2.87)	0.33 (0.54)	5.61 (2.83)	1.76 (0.89)
ADAS-Cog11	10.8 (8.0)	5.3 (2.9)	22.5 (9.3)	10.7 (2.3)
ADAS-Cog13	17.0 (10.8)	8.3 (4.5)	33.0 (10.2)	19.1 (3.0)
Mini-Mental State Examination	27.1 (3.4)	28.9 (1.2)	21.5 (4.2)	26.9 (12.3)
RAVLT immediate	35.4 (11.3)	46.0 (11.1)	20.4 (7.9)	29.7 (6.0)
RAVLT learning	4.2 (2.6)	5.7 (2.3)	1.6 (1.7)	3.0 (1.9)
RAVLT forgetting	4.2 (2.4)	3.6 (2.8)	4.2 (1.8)	4.6 (2.1)
RAVLT forgetting percent	56.0 (33.4)	34.5 (30.7)	92.8 (17.6)	63.2 (28.5)
Functional Activities Questionnaire	5.5 (8.1)	0.6 (1.5)	1.6 (7.5)	2.8 (1.7)
Montreal Cognitive Assessment	2.11 (0.42)	2.60 (0.24)	16.2 (4.8)	22.8 (1.9)
Ventricles	$4.11\ (2.03) \times 10^4$	$2.96\ (1.42) \times 10^4$	$5.21\ (2.50) \times 10^4$	$8.23\ (2.86) \times 10^4$
Hippocampus	$6.33\ (0.88) \times 10^3$	$7.65\ (0.85) \times 10^3$	$5.61\ (1.08) \times 10^3$	$5.91\ (0.30) \times 10^3$
Whole brain volume	$9.50\ (0.08) \times 10^6$	$1.05\ (0.09) \times 10^6$	$0.96\ (0.11) \times 10^6$	$1.04\ (0.02) \times 10^6$
Entorhinal cortical volume	$3.46\ (0.82) \times 10^3$	$3.97\ (0.58) \times 10^3$	$2.74\ (0.71) \times 10^3$	$3.38\ (0.24) \times 10^3$
Fusiform cortical volume	$1.61\ (0.19) \times 10^4$	$1.86\ (0.22) \times 10^4$	$1.51\ (0.27) \times 10^4$	$1.65\ (0.13) \times 10^4$

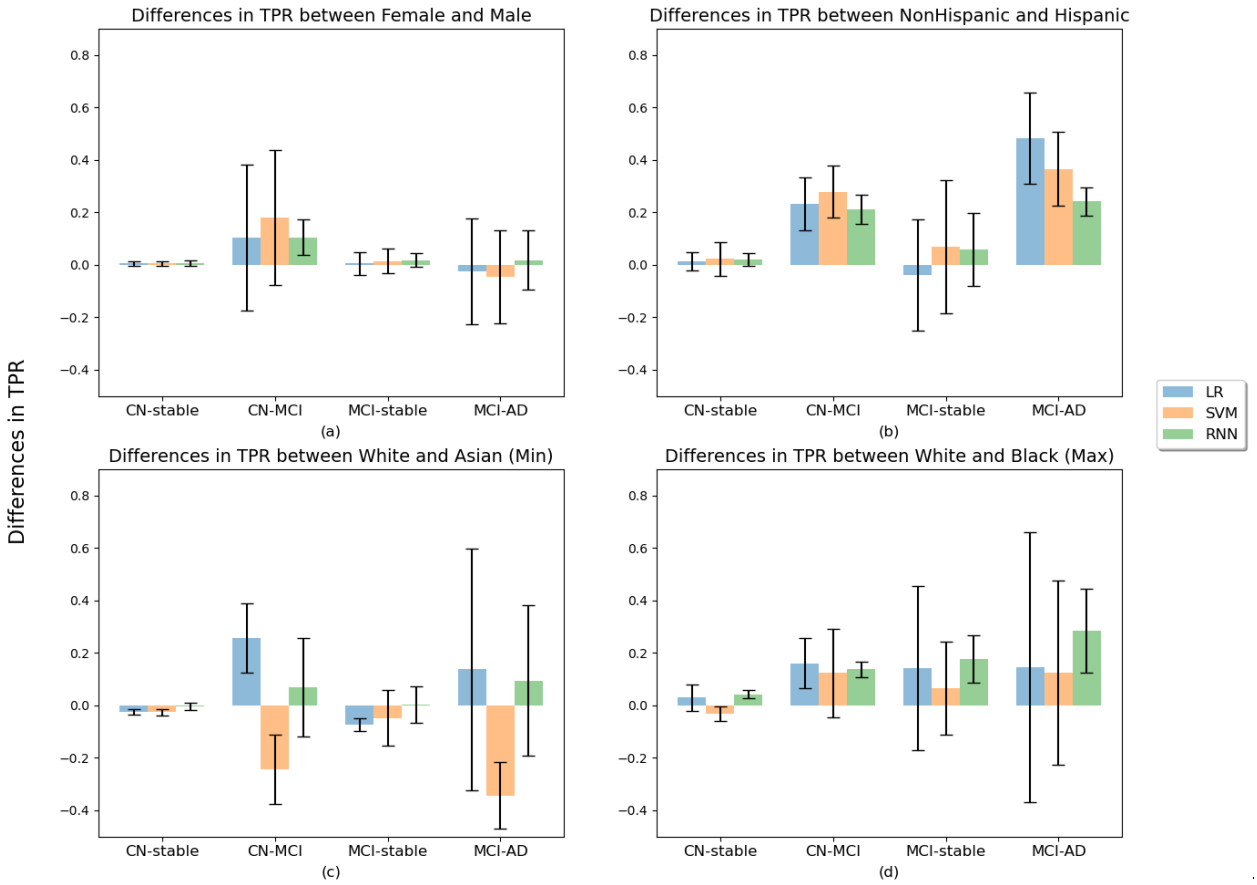
Middle temporal cortical volume	1.81 (0.27) × 10^4	2.08 (0.26) × 10^4	1.68 (0.32) × 10^4	1.91 (0.23) × 10^4
Intracranial volume	1.48 (0.15) × 10^6	1.50 (0.14) × 10^6	1.53 (0.18) × 10^6	1.67 (0.12) × 10^6
Florbetapir (18F-AV-45) - PET	1.3 (0.2)	1.1 (0.1)	1.1 (0.1)	1.3 (0.2)
Fluorodeoxyglucose (FDG) - PET	1.2 (0.1)	1.3 (0.1)	1.3 (0.1)	1.2 (0.1)
Beta-amyloid (CSF)	0.79 (0.44) × 10^3	1.40 (0.57) × 10^3	0.64 (0.38) × 10^3	0.56 (0.09) × 10^3
Total tau	3.13 (0.92) × 10^2	2.32 (0.76) × 10^2	3.69 (1.41) × 10^2	2.37 (0.09) × 10^2
Phosphorylated tau	33.1 (1.1)	20.9 (7.8)	36.5 (15.0)	22.2 (0.7)

Note: AD-stable indicates people observed with the same stage at baseline and final visit; CN-AD denotes CN progress to AD; MCI-CN denotes that MCI progress to CN; AD-MCI denotes that AD convert to MCI. SB: Sum of boxes, ADAS: Alzheimer's Disease Assessment Scale, RAVLT: Rey Auditory Verbal Learning Test.

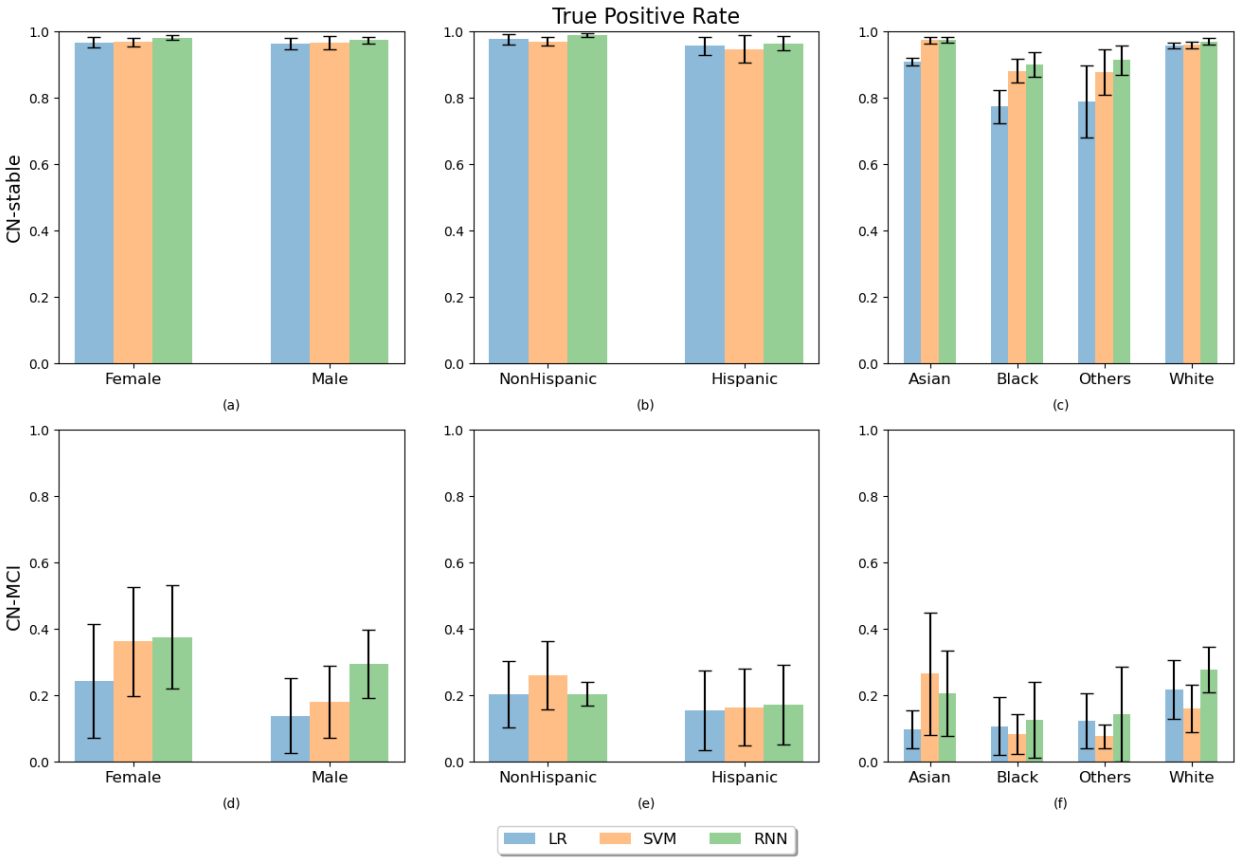
eTable 2. Prediction performance averaged across 10 test sets.

	mAUC (mean ± SD)	BCA (mean ± SD)
LR	0.916 ± 0.017	0.825 ± 0.023
SVM	0.921 ± 0.011	0.831 ± 0.021
RNN	0.949 ± 0.008	0.891 ± 0.017

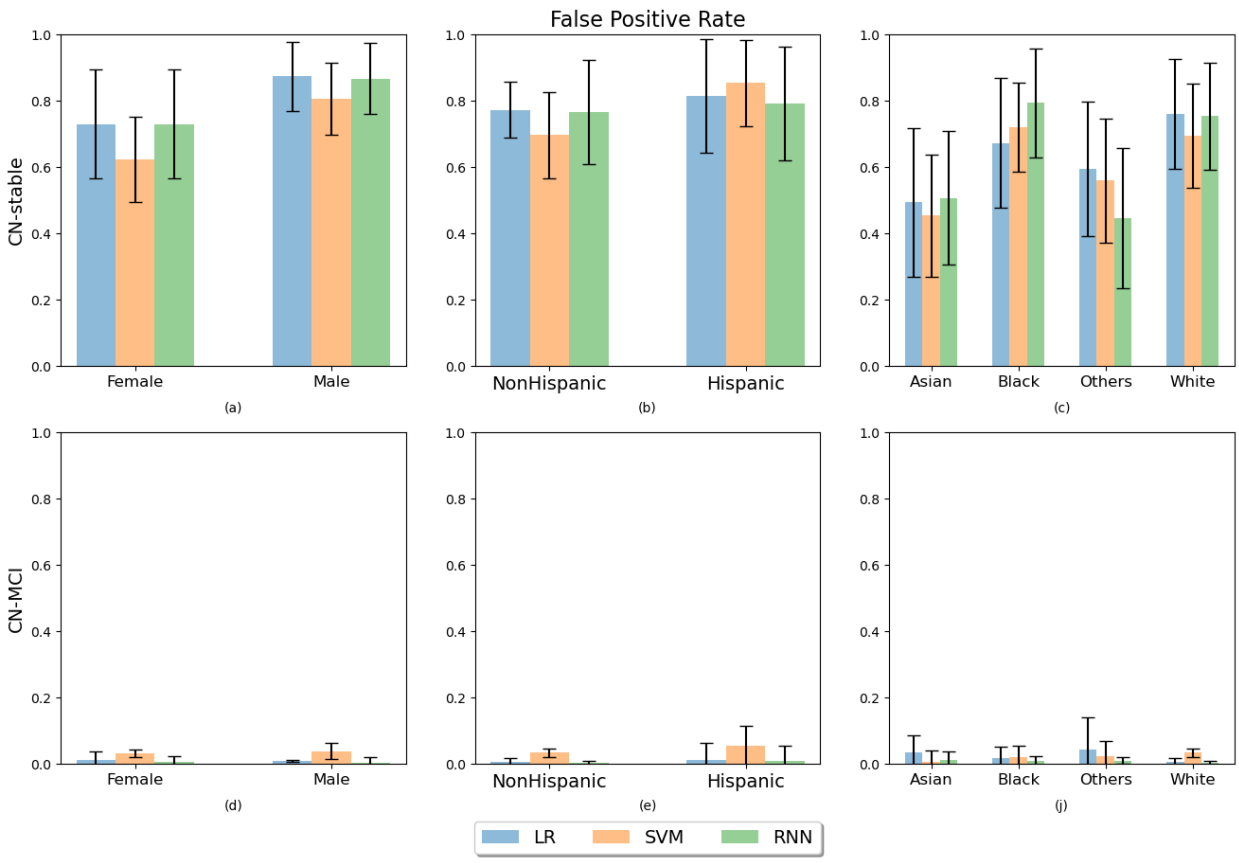
In each cell, the two numbers represent the mean and standard deviation derived from 10 tests. mAUC = multiclass area under the operating curve; BCA = balanced class accuracy. LR= logistic regression; SVM = Support Vector Machine; RNN = recurrent neural networks.



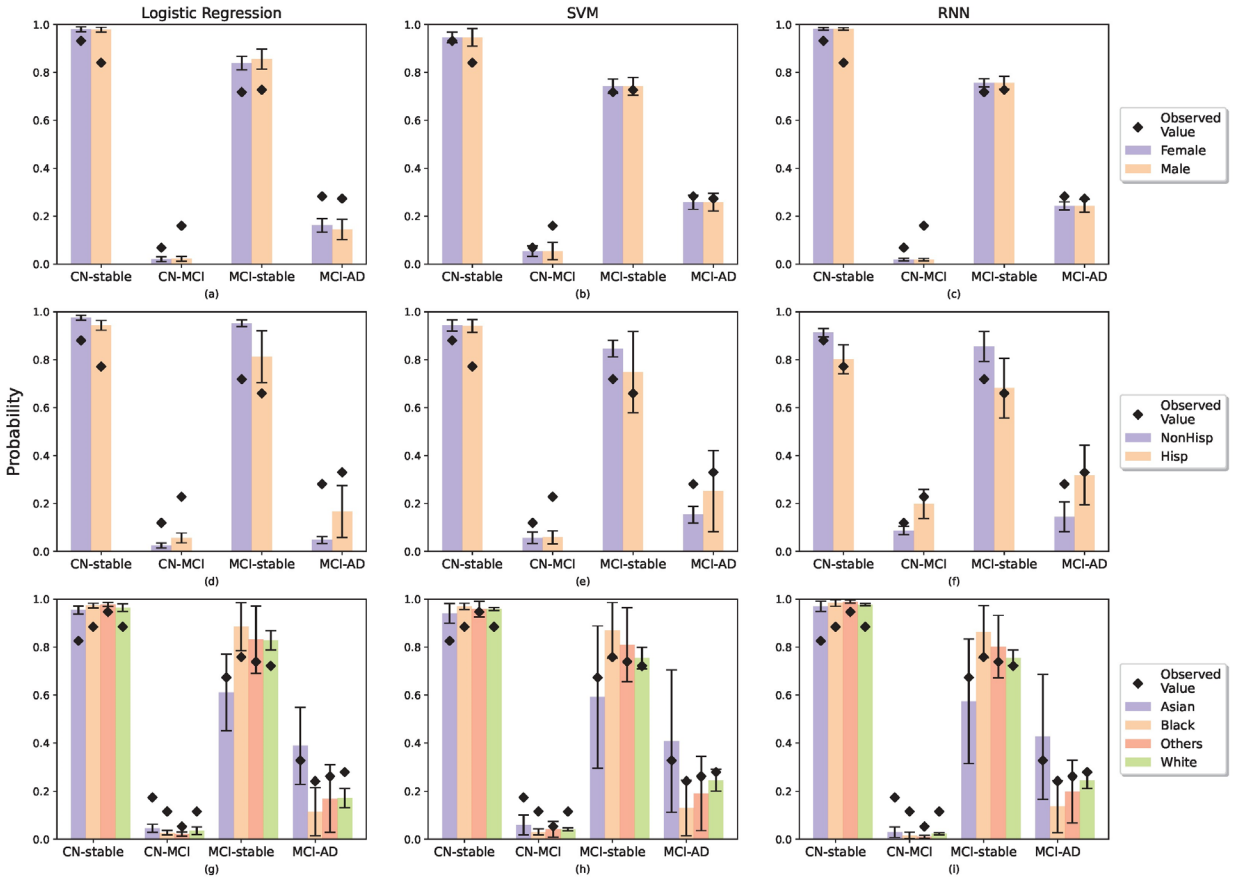
eFigure 2 Absolute differences in TPR across groups defined by the three protected attributes. For gender, the difference in TPR is between male and female. For ethnicity, the difference in TPR is reported between Non-Hispanic and Hispanic. For race, since there are four groups, we first compute all pairwise differences with the "White" group which we considered as a reference as it had the largest sample size. We then report the minimum differences in TPR, which resulted from the contrast between the White and Asian groups, and the maximum differences, which resulted from comparing the White and Black groups. Bars represent mean values across 10 test sets and error bars represent a corresponding standard deviation of the 10 mean values.



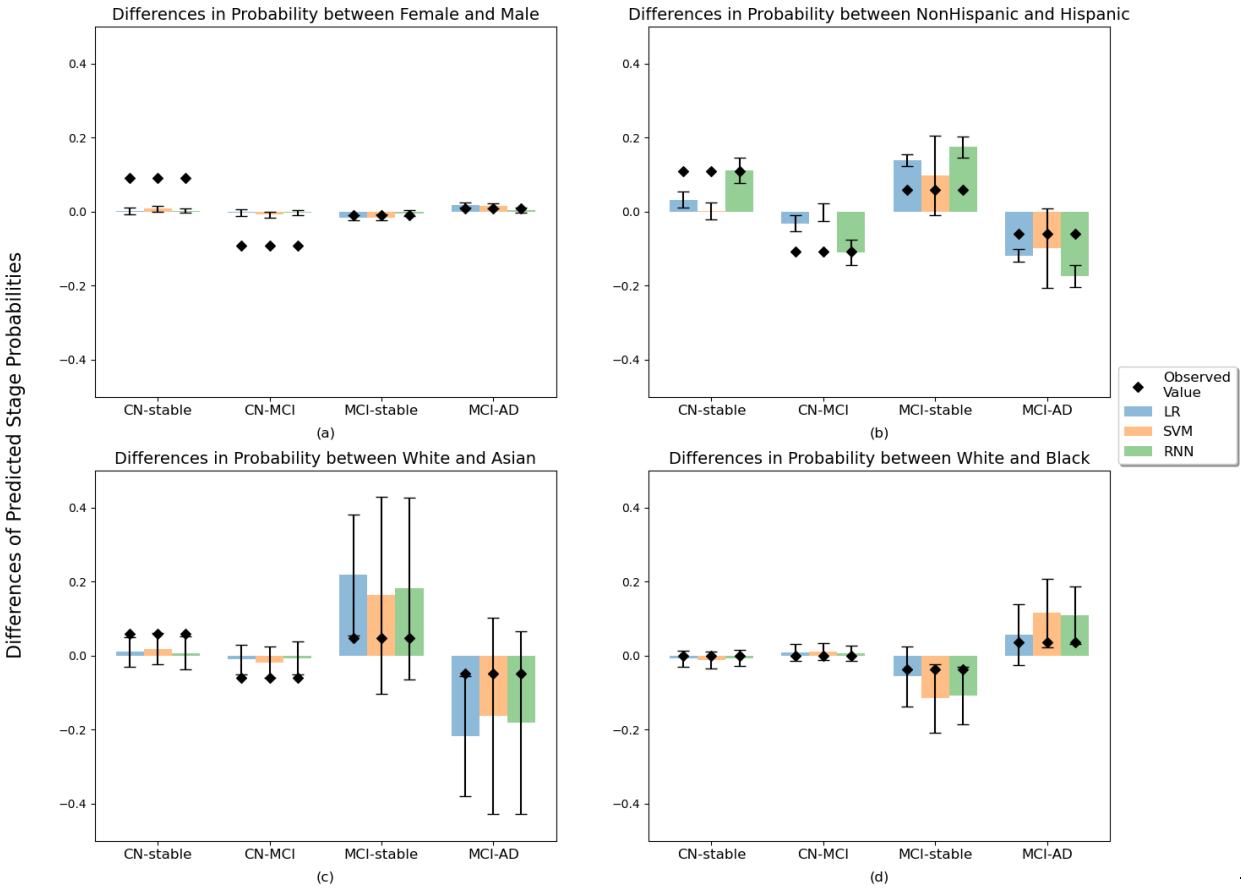
eFigure 3. Comparison of True Positive Rates across subgroups of gender, ethnicity and race for three models for participants with cognitively normal at baseline. The results are averaged over 10 test sets using predictions from the LR, SVM, and RNN models. Bars present the mean values across 10 test sets and error bars represent the standard deviation of the 10 mean values.



eFigure 4. Comparison of False Positive Rates across subgroups of gender, ethnicity and race for three models for participants with cognitively normal at baseline. The results are averaged over 10 test sets using predictions from the LR, SVM, and RNN models. Bars present the mean values across 10 test sets and error bars represent the standard deviation of the 10 mean values.



eFigure 5. Comparison of predicted probability of progression cases across subgroups of gender, ethnicity, and race for three models. The results are averaged over 10 test sets using predictions from the LR, SVM, and RNN models. Bars represent the mean values across 10 test sets and error bars represent a corresponding standard deviation of the 10 mean values. Dots represent the average value of the empirical probability of each trajectory stratified by demographic subgroup on 10 test sets.



eFigure 6. Differences of predicted progression probabilities between groups of each protected attribute with three evaluated models. The results are averaged over 10 test sets using predictions from the LR, SVM, and RNN models. Bars represent the mean values across 10 test sets and error bars represent a corresponding standard deviation of the 10 mean values. Dots represent the average values of differences of the empirical probability of each trajectory stratified by demographic subgroup on 10 test sets.