JAMA Network Open

Original Investigation | Geriatrics

Algorithmic Fairness of Machine Learning Models for Alzheimer Disease Progression

Chenxi Yuan, PhD; Kristin A. Linn, PhD; Rebecca A. Hubbard, PhD

Abstract

IMPORTANCE Predictive models using machine learning techniques have potential to improve early detection and management of Alzheimer disease (AD). However, these models potentially have biases and may perpetuate or exacerbate existing disparities.

OBJECTIVE To characterize the algorithmic fairness of longitudinal prediction models for AD progression.

DESIGN, SETTING, AND PARTICIPANTS This prognostic study investigated the algorithmic fairness of logistic regression, support vector machines, and recurrent neural networks for predicting progression to mild cognitive impairment (MCI) and AD using data from participants in the Alzheimer Disease Neuroimaging Initiative evaluated at 57 sites in the US and Canada. Participants aged 54 to 91 years who contributed data on at least 2 visits between September 2005 and May 2017 were included. Data were analyzed in October 2022.

EXPOSURES Fairness was quantified across sex, ethnicity, and race groups. Neuropsychological test scores, anatomical features from T1 magnetic resonance imaging, measures extracted from positron emission tomography, and cerebrospinal fluid biomarkers were included as predictors.

MAIN OUTCOMES AND MEASURES Outcome measures quantified fairness of prediction models (logistic regression [LR], support vector machine [SVM], and recurrent neural network [RNN] models), including equal opportunity, equalized odds, and demographic parity. Specifically, if the model exhibited equal sensitivity for all groups, it aligned with the principle of equal opportunity, indicating fairness in predictive performance.

RESULTS A total of 1730 participants in the cohort (mean [SD] age, 73.81 [6.92] years; 776 females [44.9%]; 69 Hispanic [4.0%] and 1661 non-Hispanic [96.0%]; 29 Asian [1.7%], 77 Black [4.5%], 1599 White [92.4%], and 25 other race [1.4%]) were included. Sensitivity for predicting progression to MCI and AD was lower for Hispanic participants compared with non-Hispanic participants; the difference (SD) in true positive rate ranged from 20.9% (5.5%) for the RNN model to 27.8% (9.8%) for the SVM model in MCI and 24.1% (5.4%) for the RNN model to 48.2% (17.3%) for the LR model in AD. Sensitivity was similarly lower for Black and Asian participants compared with non-Hispanic White participants; for example, the difference (SD) in AD true positive rate was 14.5% (51.6%) in the LR model, 12.3% (35.1%) in the SVM model, and 28.4% (16.8%) in the RNN model for Black vs White participants, and the difference (SD) in MCI true positive rate was 25.6% (13.1%) in the LR model, 24.3% (13.1%) in the SVM model, and 6.8% (18.7%) in the RNN model for Asian vs White participants. Models generally satisfied metrics of fairness with respect to sex, with no significant differences by group, except for cognitively normal (CN)–MCI and MCI-AD transitions (eg, an absolute increase [SD] in the true positive rate of CN-MCI transitions of 10.3% [27.8%] for the LR model).

(continued)

Key Points

Question Do algorithms to predict Alzheimer disease exhibit vulnerabilities to biases that are associated with unfair decisions and favoring of specific groups of people?

Findings In this prognostic study of 1730 participants in the Alzheimer Disease Neuroimaging Initiative, models predicting progression to mild cognitive impairment and Alzheimer disease had lower sensitivity for Hispanic participants compared with non-Hispanic participants and for Black and Asian participants compared with non-Hispanic White participants. Models generally satisfied metrics of fairness with respect to sex.

Meaning This study found that machine learning models had notable deficits in fairness across race and ethnicity groups.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Open Access. This is an open access article distributed under the terms of the CC-BY-NC-ND License.

Downloaded from jamanetwork.com by guest on 01/15/2024

Abstract (continued)

CONCLUSIONS AND RELEVANCE In this study, models were accurate in aggregate but failed to satisfy fairness metrics. These findings suggest that fairness should be considered in the development and use of machine learning models for AD progression.

JAMA Network Open. 2023;6(11):e2342203. doi:10.1001/jamanetworkopen.2023.42203

Introduction

The development and use of machine learning (ML) algorithms in health care has received a surge of attention in recent years.[1-12] Although ML algorithms can inform clinical decision-making and are potentially associated with improved population health,[13] there is growing concern that ML may inadvertently introduce bias into decision-making processes, which may be associated with unintended discrimination against underrepresented and disadvantaged populations.[14-16] Given that algorithms are vulnerable to biases that render their decisions unfair, *fairness*, in the context of decision making, is the absence of any prejudice or favoritism toward an individual or group based on that groups' inherent or acquired characteristics. An unfair algorithm, also referred to as algorithmic bias, skews benefits toward a particular group of people with respect to protected attributes.[17] Protected attributes are features that may not be used as the basis for decisions. There is no one universal set of protected attributes. They are determined based on laws, regulations, or other policies governing a particular application domain in a particular jurisdiction. Attributes such as race and ethnicity, color, age, gender and sex, national origin, religion, and marital status are commonly considered protected attributes.[17-19]

The Department of Health and Human Services has mandated identification of sources of bias and discriminatory outputs in ML algorithms,[20] and a large body of research has been conducted on algorithmic bias in health and medicine.[21-27] However, the problem of algorithmic bias in the context of ML for Alzheimer disease (AD), such as the prediction of AD progression using ML approaches, has received little attention. Biased prediction models may favor or disadvantage some groups, which may be associated with misdiagnoses, improper treatment recommendations, and insufficient or unnecessary care for individuals experiencing bias.[28-30] In this study, we investigated the algorithmic fairness of longitudinal prediction models for AD progression. Using publicly available data from the Alzheimer Disease Neuroimaging Initiative (ADNI),[31] we had an objective of auditing the fairness of ML models for AD progression prediction. The overall goals of this study were to introduce and define fairness metrics relevant to models for predicting AD progression and to illustrate how ML algorithms may be analyzed to reveal potential disparities across protected attributes.

Methods

This prognostic study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline. Written informed consent was obtained for participation in ADNI, as approved by the institutional review board at each participating center. Ethical approval and informed consent were not required because this study, consisting of secondary data analysis of coded data that cannot be linked to individual participants, is not considered human participants research according to the policies of the Institutional Review Board of the University Pennsylvania.

Population

Data were derived from the ADNI to facilitate study of AD progression.[31,32] In brief, ADNI enrolled participants aged 54 to 91 years at 57 sites in the US and Canada. Our data set incorporated longitudinal data from multiple ADNI study phases and measurements from every participant

Downloaded from jamanetwork.com by guest on 01/15/2024

contributing data on at least 2 visits between September 2005 and May 2017. Clinical status at each visit was classified as cognitively normal (CN), mild cognitive impairment (MCI), or AD.

Protected Attributes

To evaluate fairness criteria, groups were defined by demographic attributes. We focused on attributes of sex, ethnicity, and race because previous studies in the fairness literature have highlighted algorithmic bias according to these characteristics.[27,33] All characteristics were classified according to participant self-report. Sex was classified as female or male. Ethnicity was classified as not Hispanic or Latino or Hispanic or Latino. Participants reporting unknown ethnicity were excluded from ethnicity-stratified analyses. Race included 7 distinct groups: Asian, American Indian or Alaskan Native, Black or African American, Hawaiian or Other Pacific Islander, White, and more than 1 reported race. We aggregated America Indian and Alaskan Native, Hawaiian and Other Pacific Islander, more than 1 race, and unknown into a category labeled other due to their small population sizes and evaluated fairness across 4 racial categories: Asian, Black, White, and other.

Study Design

We defined unfairness, or algorithmic bias, as differences in the predictive performance of an ML algorithm across subpopulations defined by a protected attribute. For example, differences in sensitivity of a model for predicting AD progression in a Black population compared with a White population would be indicative of unfairness. We focused on 3 commonly used fairness metrics: equalized odds, equal opportunity, and demographic parity.[17,34] These criteria have natural interpretations in the context of AD progression prediction. Equal opportunity is defined as equal sensitivity or true positive rates (TPRs) of the ML algorithm across all levels of the protected attribute.[35] An AD progression algorithm would exhibit equal opportunity if individuals who progressed to AD were equally likely to be identified by the algorithm across all protected groups. Equalized odds requires that an algorithm exhibit equal opportunity and equal specificity or false positive rates (FPRs) across groups. Demographic parity is the equivalence of a predicted event's probability across groups by sensitive attribute.[34] For progression to AD, demographic parity with respect to sex would be satisfied if females and males were predicted to develop AD with equal probability. When real differences in outcome prevalence exist across groups, achieving demographic parity may be undesirable. Importantly, unless prevalence is equal across groups, it is impossible to simultaneously satisfy all metrics. **Table 1** presents mathematical definitions of these fairness metrics.

Prediction Models

We assessed fairness with respect to the task of predicting AD progression with ML algorithms.[36] We selected 3 ML models for evaluation in this study: logistic regression (LR), support vector machine (SVM), and recurrent neural network (RNN) models. We included LR and SVM because they are well-established ML models commonly used for prediction problems and are often presented as comparators for new models.[7,37,38] As a deep-learning model, RNN has shown promise in the AD progression domain[39,40] and has been applied to prediction problems,[38,41] demonstrating improvement over other ML models on prediction accuracy. The RNN model we tested in this study

Table 1. Mathematical Definitions of 3 Common Fairness Metrics

Fairness metric	Definition[a]	Explanation with sample of use in Alzheimer disease
Equal opportunity	True positive rates are the same across groups: $P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1)$	The probability of correctly predicting that an individual progresses to AD is the same for groups defined by a protected attribute, such as race
Equal odds	True positive rates and false positive rates are the same across groups: $P(\hat{Y} = 1 \mid A = 0, Y = y) = P(\hat{Y} = 1 \mid A = 1, Y = y), y \in \{0,1\}$	The probability of correctly predicting that an individual progresses to AD and the probability of incorrectly predicting progression to AD for those who do not are the same for groups defined by a protected attribute, such as race
Demographic parity	Equal probability of being classified with the positive label: $P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$	The proportion of individuals predicted to progress to AD is the same across groups defined by a protected attribute, such as race

[a] Y indicates the observed outcome, $\hat{Y}$ a prediction of Y, and A a binary protected attribute.

Downloaded from jamanetwork.com by guest on 01/15/2024

is from Nguyen et al.[36] Given multimodal predictors and the diagnostic status of a participant at baseline, we sought to predict diagnosis stage at the subsequent visit as CN, MCI, or AD for all subsequent months. We defined progression trajectories as transition from baseline CN to MCI, baseline MCI to AD, stable CN, and stable MCI (ie, patients recorded at the same stage at baseline and at a final visit). Predictors included neuropsychological test scores, anatomical features derived from magnetic resonance imaging, positron emission tomography measures, and cerebrospinal fluid markers (see complete list of predictors in **Table 2**). Additional details of model implementation and training are provided in eMethods in Supplement 1.

We used cross-validation for model selection and evaluation. Data were randomly partitioned into 10 equal subsets. Each 10-fold cross-validation iteration used 80% of participants for training, 10% of participants for model validation, and 10% of participants for testing. The training set was used for model fitting, the validation set for hyperparameter selection, and the test set for model performance evaluation. All continuous variables were *z* normalized using the training set to estimate the mean and SD, which were then used to *z* normalize validation and test sets. The multiclass area under the operating curve and balanced class accuracy were used to evaluate models (eMethods and eTable 2 in Supplement 1).

Statistical Analysis

To assess algorithmic fairness, we calculated fairness metrics on each of 10 test sets using predictions from each model. Metrics are reported as the mean and SD across the 10 values. Evaluations were conducted separately by demographic group. We first assessed equal opportunity by computing the TPR for groups defined by each protected attribute separately for each cognitive functioning trajectory (ie, CN to MCI, MCI to AD, stable CN, and stable MCI) and each of 3 models. The TPR quantifies the proportion of individuals experiencing a given trajectory who were correctly predicted to follow that trajectory. For instance, the TPR of CN to MCI represents the probability of correctly predicting that an individual progressed from CN to MCI. A TPR value of 1 indicates that the model has achieved perfect sensitivity in identifying the positive instances within the category. If the TPRs for each trajectory are similar across protected feature categories, it suggests that the model attained equal opportunity. We also calculated differences in TPR between groups for each protected attribute. In addition to TPR, we calculated the FPR. Specifically, for a given trajectory the FPR is defined as the proportion of individuals who did not experience that trajectory who were incorrectly predicted to follow that trajectory. For example, the FPR of CN to MCI represents the probability of predicting progression from CN to MCI for an individual who did not progress. An algorithm must demonstrate equal TPR and equal FPR across groups to satisfy the equalized odds criterion. To assess demographic parity, we computed the predicted probability for each individual. We report the difference in mean predicted probabilities across groups for each sensitive attribute, trajectory, and ML model. Finally, we calculated the empirical probability of each trajectory stratified by demographic group. Hypothesis testing was not used in this context due to the lack of independence among predicted values on test sets.[42] Additionally, due to small sample sizes in some groups, nonparametric and parametric tests are expected to have low power. We therefore focused on point estimation and interpretation of point estimates given their variability. An overview of the experimental procedure is shown in eFigure 1 in Supplement 1. Statical analysis was conducted using Python programming language version 2.7 (Python Software Foundation). Data were analyzed in October 2022.

Results

Study Cohort

A total of 1730 participants (mean [SD] age, 73.81 [6.92] years; 776 females [44.9%]; 69 Hispanic [4.0%] and 1661 non-Hispanic [96.0%]; 29 Asian [1.7%], 77 Black [4.5%], 1599 White [92.4%], and 25 other race [1.4%]) were included, and each was scanned at multiple time points, contributing an

Downloaded from jamanetwork.com by guest on 01/15/2024

mean (SD) 7.3 (4.0) observations per participant over a mean [SD] 3.6 (2.5) years. The distribution of participant characteristics stratified by clinical status at the baseline and last visit is provided in Table 2. There were 337 participants with the CN-stable trajectory (86.2%), 54 participants with the CN-MCI trajectory (13.8%), 519 participants with the MCI-stable trajectory (62.4%), and 313 individuals with the MCI-AD trajectory (38.6%). These groups included 173 females (51.3%), 19

Table 2. Participant Characteristics by Cognitive Functioning Trajectory

	Mean (SD) (N = 1730)[a]			
Characteristic	CN-stable trajectory (n = 337 [86.2%])	CN-MCI trajectory (n = 54 [13.8%])	MCI-stable trajectory (n = 519 [62.4%])	MCI-AD trajectory (n = 313 [37.6%])
Protected attribute, No. (%)				
Sex				
Female	173 (51.3)	19 (35.2)	213 (41.0)	123 (39.3)
Male	164 (48.7)	35 (64.8)	306 (59.0)	190 (60.7)
Ethnicity				
Hispanic	13 (3.8)	5 (9.3)	20 (3.9)	10 (3.2)
Not Hispanic	324 (96.2)	49 (90.7)	499 (96.1)	303 (96.8)
Race				
Asian	7 (2.1)	2 (3.7)	7 (1.3)	6 (2.0)
Black	24 (7.1)	5 (9.3)	22 (4.2)	7 (2.2)
White	303 (89.9)	42 (77.7)	479 (92.3)	298 (95.2)
Other[b]	3 (0.9)	5 (9.3)	11 (2.2)	2 (0.6)
Predictor				
CDR-SB	0.08 (0.46)	0.45 (0.77)	1.41 (1.21)	4.11 (3.28)
ADAS-Cog				
11	5.4 (0.2)	7.3 (3.7)	9.0 (4.7)	16.7 (9.0)
13	84.9 (4.3)	11.7 (5.5)	14.4 (6.9)	25.8 (11.1)
Mini Mental State Examination	29.0 (1.2)	28.8 (1.4)	27.7 (2.2)	24.3 (4.5)
RAVLT				
Immediate	45.4 (10.4)	39.4 (10.6)	35.7 (11.3)	25.2 (9.2)
Learning	5.8 (2.4)	4.8 (2.5)	4.3 (2.6)	2.4 (2.2)
Forgetting	3.4 (2.8)	4.2 (2.4)	4.5 (2.5)	4.7 (2.1)
RAVLT Percent Forgetting	32.5 (31.9)	47.6 (30.0)	56.7 (36.6)	83.3 (30.4)
Functional Activities Questionnaire	1.8 (8.2)	0.9 (2.2)	2.6 (3.9)	1.1 (0.8)
Montreal Cognitive Assessment	25.8 (2.5)	24.4 (2.8)	23.8 (3.1)	18.6 (5.3)
Volume, mm^3				
Ventricles, $\times 10^4$	3.50 (1.95)	4.23 (1.93)	4.01 (2.32)	4.88 (2.37)
Hippocampus, $\times 10^3$	7.32 (0.92)	6.89 (0.86)	6.97 (1.11)	5.91 (1.11)
Whole brain, $\times 10^6$	1.01 (0.10)	1.02 (0.09)	1.04 (0.10)	0.98 (0.11)
Entorhinal cortical, $\times 10^3$	3.79 (0.61)	3.56 (0.76)	3.64 (0.71)	2.99 (0.78)
Fusiform cortical, $\times 10^4$	1.76 (0.24)	1.76 (0.23)	1.80 (0.26)	1.59 (0.27)
Middle temporal cortical, $\times 10^4$	2.00 (0.26)	1.97 (0.24)	2.01 (0.27)	1.77 (0.30)
Intracranial, $\times 10^6$	1.51 (0.15)	1.56 (0.14)	1.53 (0.16)	1.54 (0.17)
PET				
18F-AV-45	1.0 (0.1)	1.1 (0.1)	1.1 (0.2)	1.3 (0.2)
FDG	1.3 (0.1)	1.2 (0.1)	1.3 (0.1)	1.1 (0.1)
CSF-β-amyloid level, pg/mL $\times 10^3$	1.31 (0.61)	1.31 (0.75)	1.11 (0.58)	0.68 (0.31)
τ level, pg/mL				
Total, $\times 10^2$	2.40 (0.90)	2.87 (0.87)	2.69 (1.18)	3.50 (1.46)
Phosphorylated	22.0 (9.5)	26.6 (8.5)	25.5 (13.2)	34.6 (16.3)

Abbreviations: 18F-AV-45, florbetapir; AD, Alzheimer disease; ADAS-Cog, Alzheimer Disease Assessment Scale-Cognitive Subscale; CDR-SB, Clinical Dementia Rating-Sum of Boxes; CN, cognitively normal; FDG, fluorodeoxyglucose; MCI, mild cognitive impairment; PET, positron emission tomography; RAVLT, Rey Auditory Verbal Learning Test.

[a] In the CN-MCI trajectory, participants had CN progression to MCI. In the MCI-AD trajectory, participants had MCI progression to AD. In CN-stable and MCI-stable trajectories, participants were observed with the same stage at baseline and final visit.

[b] Includes American Indian or Alaskan Native, Hawaiian or Other Pacific Islander, and more than 1 reported race, and unknown race.

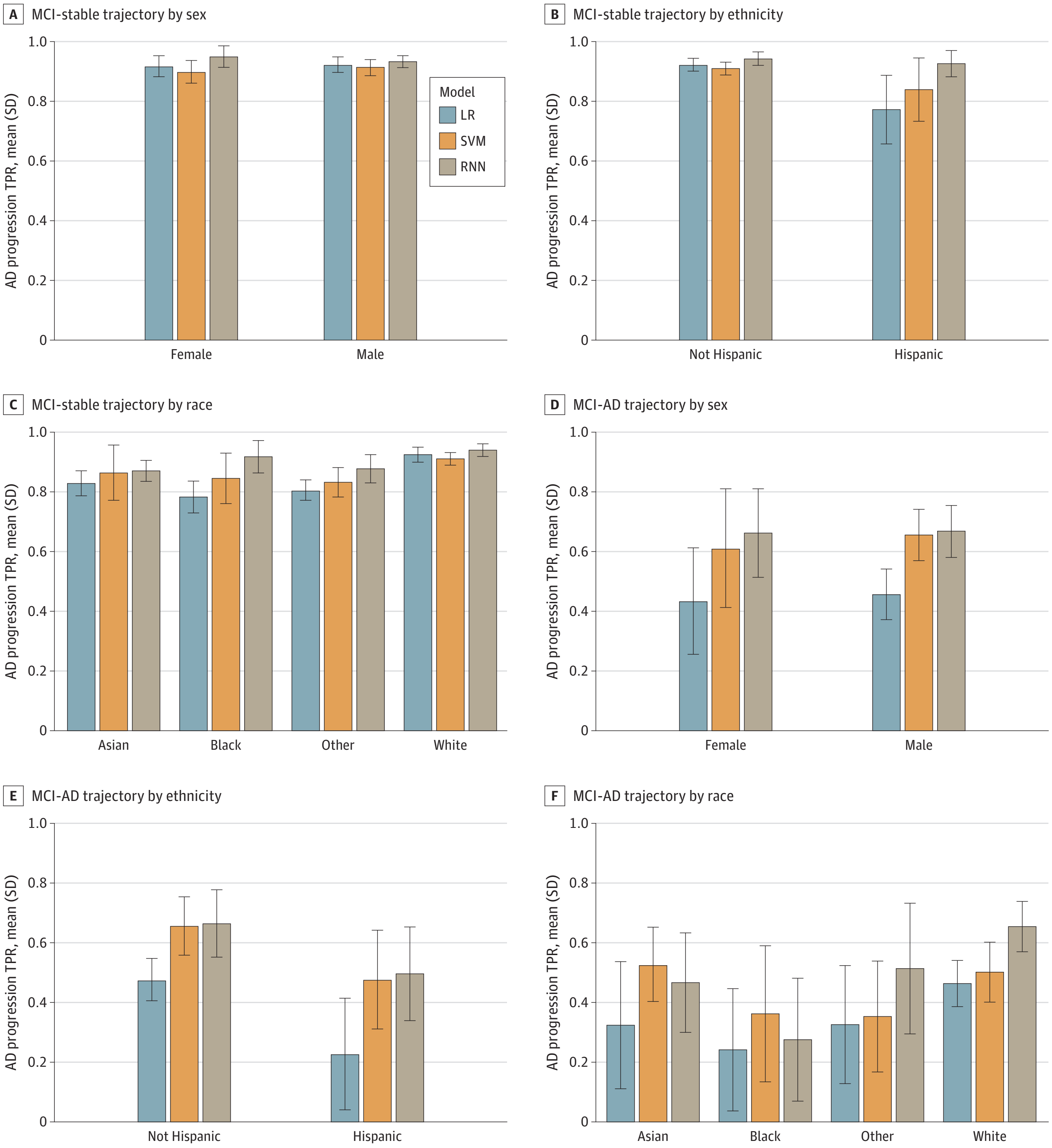
Downloaded from jamanetwork.com by guest on 01/15/2024

females (35.2%), 213 females (41.0%), and 123 females (39.3%), respectively. Backward transitions (ie, MCI to CN or AD to MCI or CN) and transitions from CN to AD were rarely observed (eTable 1 in Supplement 1) and were, therefore, not included in fairness evaluations.

Equal Opportunity and Equalized Odds

Figure 1A and D and eFigure 3 in Supplement 1 show TPRs for progression to AD and MCI, respectively, stratified by sex for each of 3 models. For CN-stable and MCI-stable trajectories, the TPR

Figure 1. True Positive Rates (TPRs) of Alzheimer Disease Progression by Protected Attribute



TPRs across sex, ethnicity, and race groups are compared for 3 models for participants with mild cognitive impairment (MCI) at baseline. Results are given for the mean over 10 test sets using predictions from logistic regression (LR), support vector machine (SVM), and recurrent neural network (RNN) models. Bars indicate mean values across 10 test sets; error bars, SDs of 10 mean values.

Downloaded from jamanetwork.com by guest on 01/15/2024

was close to 1, and there were no major differences in TPR between females and males. The difference (SD) in TPR between males and females for the CN-stable trajectory was 0.5% (0.8%) for the LR model, 0.5% (0.8%) for the SVM model, and 0.6% (0.9%) for the RNN model. The difference (SD) in TPR for the MCI-stable trajectory was 0.4% (4.3%) for the LR model, 1.3% (4.6%) for the SVM model, and 1.7% (2.7%) for the RNN model (eFigure 2 in Supplement 1). For transition from CN to MCI, there was a notable difference in TPR between sexes, with all models performing better for females than males; there was an absolute increases (SD) of 10.3% (27.8%), 15.0% (25.7%), and 10.4% (6.8%) for LR, SVM, and RNN models, respectively. For transition from MCI to AD, small differences in TPR were observed between sexes (eg, the difference [SD] was 2.5% [20.1%] in the LR model, 4.6% [17.5%] in the SVM model, and 1.7% [11.2%] in the RNN model). Models performed similarly overall, but RNN had higher TPR for predicting progression from CN to MCI and MCI to AD, as well as less variability across test sets (eFigure 2 in Supplement 1).

Figure 1B and E and eFigure 3 in Supplement 1 show TPRs for progression to AD and MCI, respectively, stratified by ethnicity. Overall, across trajectories and models, TPR was higher for non-Hispanic participants compared with Hispanic participants. The difference (SD) in TPR between non-Hispanic and Hispanic participants was 1.4% (3.5%) in the LR model, 2.2% (6.5%) in the SVM model, and 2.1% (6.5%) in the RNN model for the CN-stable trajectory and 3.9% (21.3%) in the LR model, 6.9% (25.3%) in the SVM model, and 5.8% (25.3%) in the RNN model for the MCI-stable trajectory across the 3 models (eFigure 2 in Supplement 1). Differences in TPRs were larger for progression from CN to MCI and MCI to AD. Specifically, the TPR difference (SD) for Hispanic participants was 23.1% (10.1%), 27.8% (9.8%), 20.9% (5.5%) lower compared with non-Hispanic participants for progression from CN to MCI and 48.2% (17.3%), 36.6% (13.9%), 24.1% (5.4%) lower for progression from MCI to AD for LR, SVM, and RNN models, respectively. In most cases, RNN had a higher TPR than other models. Across models for MCI progression to AD, RNN had the highest TPR and smallest difference in TPR between Hispanic and non-Hispanic participants (eFigure 2 in Supplement 1).

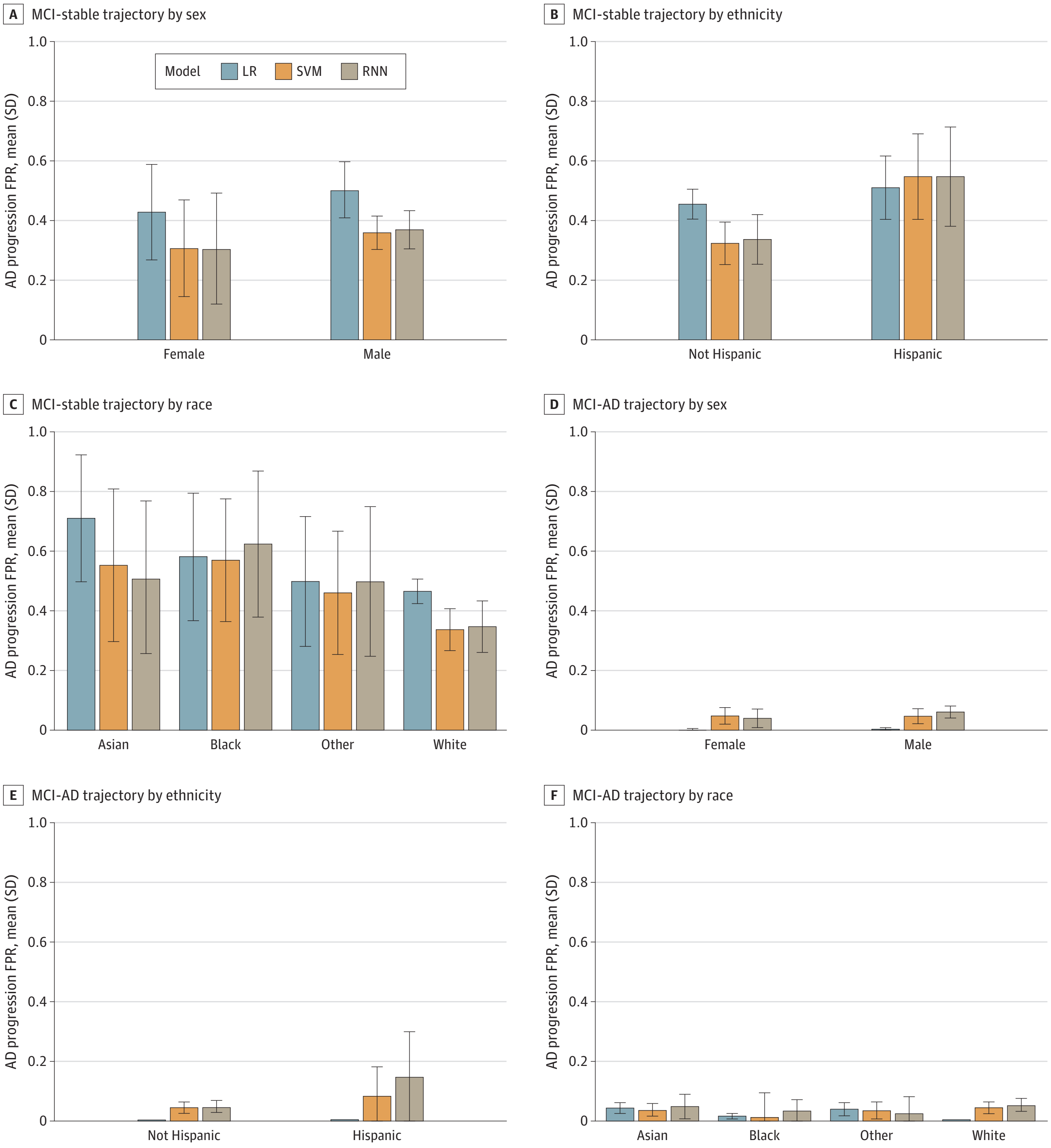
Comparisons of TPRs across racial groups are shown in Figure 1C and F for progression to AD and eFigure 3 in Supplement 1 for progression to MCI. For the CN-stable trajectory, the TPR (SD) was high for White participants (95.7% [0.9%] for the LR model, 95.9% [1.0%] for the SVM model, and 97.0% [1.1%] for the RNN model) and lower for other groups (Asian: range, 90.9% [1.1%] for the LR model to 97.3% [0.7%] for the RNN model; Black: range, 73.3% [5.1%] for the LR model to 90.0% [3.6%] for the RNN model; and other race: range, 78.9% [1.0%] for the LR model to 91.4% [4.4%] for the RNN model) across 3 models. Patterns across racial groups for the MCI-stable trajectory were similar to those for the CN-stable trajectory. For CN to MCI, Asian participants had a higher TPR (SD) than other groups for SVM (Asian: 26.5% [18.4%]; Black: 8.3% [5.9%]; White: 1.6% [7.1%]; other race: 7.6% [3.6%]). The TPR (SD) for Black participants was lowest for CN to MCI progression across 3 models (Black: range, 8.3% [5.1%] for the SVM model to 12.5% [11.4%] for the RNN model; Asian: range, 9.6% [5.7%] for the LR model to 20.6% [12.8%] for the RNN model; White: range, 16.0% [7.1%] for the SVM model to 27.7% [6.9%] for the RNN model; other race: range 7.6% [3.6%] for the LR model to 14.4% [14.3%] for the RNN model). White participants had a higher TPR for progression from MCI to AD for 2 of 3 models (eFigure 2 in Supplement 1). Sensitivity was lower for Black and Asian participants compared with non-Hispanic White participants (eg, the difference [SD] in TPR was 14.5% [51.6%] in the LR model, 12.3% [35.1%] in the SVM model, and 28.4% [16.8%] in the RNN model for AD in Black vs White participants, and the difference [SD] in TPR was 25.6% [13.1%] in the LR model, 24.3% [13.1%] in the SVM model, and 6.8% [18.7%] in the RNN model for MCI in Asian vs White participants).

For all 3 models, the FPR was lower for females compared with males for AD progression (**Figure 2**A and D) and MCI progression (eFigure 4 in Supplement 1). Similarly, non-Hispanic participants had a lower FPR than Hispanic participants for all trajectories (Figure 2B and E; eFigure 4 in Supplement 1). For racial groups, the FPR for Black participants was higher compared with that of other racial groups for the CN-stable trajectory. For the MCI-stable trajectory, the FPR was higher for

Downloaded from jamanetwork.com by guest on 01/15/2024

Asian participants compared with other groups (Figure 2C and F; eFigure 4 in Supplement 1). Overall, the FPR for progression from CN to MCI and MCI to AD was lower than that for stable CN and MCI trajectories. However, large error bars for Asian, Black, and other racial groups reflect uncertainty in FPR point estimates due to the small sample sizes of these groups.

Figure 2. False Positive Rates (FPRs) of Alzheimer Disease Progression by Protected Attribute



FPRs across sex, ethnicity, and race groups are compared for 3 models for participants with mild cognitive impairment (MCI) at baseline. Results are given for the mean over 10 test sets using predictions from logistic regression (LR), support vector machine (SVM), and recurrent neural network (RNN) models. Bars indicate mean values across 10 test sets; error bars, SDs of 10 mean values.

Downloaded from jamanetwork.com by guest on 01/15/2024

Demographic Parity

Observed and predicted prevalence of cognitive functioning trajectories differed across groups defined by protected attributes (eFigure 5 in Supplement 1). Across models, the probability of being predicted to have CN-stable or MCI-stable trajectories was higher than the observed prevalence, whereas the probability of being predicted to transition from CN to MCI or MCI to AD was generally lower than or similar to the observed prevalence.

Female participants who were CN at baseline had a higher predicted probability of a CN-stable trajectory (eFigure 5 in Supplement 1), with the difference (SD) ranging across 3 models from 0.2% (0.9%) for the LR model to 0.7% (0.8%) for the RNN model. Conversely, the predicted probabilities of MCI-stable and MCI-AD trajectories were lower for female compared with male participants, with the difference (SD) ranging from 0.4% (0.7%) for the RNN model in the MCI-stable trajectory and 0.4% (0.7%) for the RNN model in the MCI-AD trajectory to 1.6% (0.7%) for the LR model in the MCI-AD trajectory. These were similar to empirical differences in prevalence between male and female participants. Notable differences between predicted and empirical probabilities were found for male participants who were CN at baseline. Specifically, the difference (SD) between predicted and observed probabilities of progressing to MCI was 13.8% (1.0%), 9.8% (3.6%), and 13.9% (5.0%) for male participants for LR, SVM and RNN models, respectively, while for female participants, the difference (SD) was 4.8% (1.0%), 1.5% (2.1%), and 5.0% (0.6%) for LR, SVM and RNN, respectively. Across models, predictions based on the RNN model were more similar to empirical probabilities of MCI progression compared with predictions from LR and SVM models (eFigure 6 in Supplement 1).

Predicted CN-stable and MCI-stable trajectories were higher for non-Hispanic participants compared with Hispanic participants, consistent with the empirical distribution (eFigure 5 in Supplement 1). Conversely, the predicted probability of progression from CN to MCI and MCI to AD for non-Hispanic participants was lower than for Hispanic participants. The difference (SD) was 3.1% (2.1%), 0.2% (2.3%), and 13.1% (3.4%) for CN progression and 11.8% (1.6%), 9.8% (10.7%), and 17.4% (2.9%) for MCI progression across LR, SVM, and RNN models, respectively. The discrepancy (SD) between predicted and observed probabilities for Hispanic participants was 17.2% (2.0%) and 31.3% (10.7%) for the LR model, 16.9% (2.7%) and 8.8% (16.9%) for the SVM model, and 3.1% (6.0%) and 2.1% (12.4%) for the RNN model for CN progression and MCI progression, respectively.

Across racial groups, Asian participants had the lowest predicted probability of CN-stable and MCI-stable trajectories and the highest predicted probability of progression from CN to MCI and MCI to AD. Black participants had the highest predicted probability of the MCI-stable trajectory and the lowest predicted probability of progression from CN to MCI and MCI to AD (eFigure 5 in Supplement 1). Additionally, for CN-stable and CN-MCI trajectories, the largest differences between predicted and observed values were for Asian participants, with a difference (SD) of 12.8% (1.6%), 11.5% (4.1%), and 14.4% (2.1%) higher (for the CN-stable trajectory) or lower (for the CN-MCI trajectory) predicted values for LR, SVM, and RNN models, respectively (eFigure 5 in Supplement 1). For MCI-stable and MCI-AD trajectories, Black participants had the largest difference (SD) between predicted and observed values (12.6% [1.0%], 11.2% [1.3%], and 10.5% [1.3%] higher [for the MCI-stable trajectory] or lower [for the MCI-AD] trajectory for LR, SVM, and RNN models, respectively) (eFigure 5 in Supplement 1). These results indicate that Black participants with MCI at baseline were more likely to be misclassified as progressing to AD. In contrast, Asian participants who were CN at baseline were most likely to be misclassified as not progressing.

Discussion

In this prognostic study, we evaluated the fairness of ML models for predicting progression of AD across groups defined by sex, race, and ethnicity. Due to differences in prevalence of progression for males and females, ML models investigated did not satisfy the criterion of demographic parity (equal predicted probability of progression) with respect to sex. All 3 models underpredicted the probability of progressing from CN to MCI for male and female participants, but discrepancies between observed

Downloaded from jamanetwork.com by guest on 01/15/2024

and predicted probabilities of progression were larger for male participants. This finding could be attributable to greater heterogeneity of trajectories in males compared with females. Progression from MCI to AD was also underpredicted by all models. However, this underprediction was less severe for RNN compared with the other 2 models.

Models displayed unfairness with respect to multiple metrics across ethnicity groups. However, uncertainty in estimates of TPR was high for Hispanic participants due to small sample sizes, making it difficult to draw firm conclusions regarding model performance for this group. Notable discrepancies between observed and predicted probabilities of transition from CN to MCI were observed for Hispanic participants. These results highlight how underrepresentation may introduce unfairness. In the ADNI data set, 4.0% of participants were Hispanic, and consequently, models tended to perform poorly for this group. However, the deep-learning (RNN) model demonstrated improved performance relative to the other 2 models through smaller differences between predicted and observed probabilities of progression for Hispanic participants.

Estimates of model performance for participants in Asian, Black, and other race groups had wide error bars due to limited sample sizes, especially in the 2 forward-transition cases (CN to MCI and MCI to AD), making it difficult to draw conclusions. Thus, assessment of equal odds is limited for these groups. Black participants in the MCI group at baseline tended to be incorrectly predicted to transition to AD. Asian participants who were CN tended to be incorrectly underpredicted to transition to MCI. A comparison of the 3 ML models demonstrated some improvement of the deep-learning (RNN) model compared with other models. Notably, for individuals progressing to AD and Black participants, RNN outperformed other models in that discrepancies between predicted probability and observed prevalence of AD were smaller.

Sources of unfairness in ML models include sampling bias and implicit cultural biases that are reflected in the data. The health domain may also feature systemic biases inherent in biological processes that it may not be possible to mitigate.[43] In the AD domain, there are neuropsychiatric differences across racial and ethnic groups, some of which exist due to systemic racism, that are associated with disease prevalence.[25,44,45] Therefore, demographic parity may not be desirable when real differences in AD disease prevalence exist. A feasible approach to evaluating fairness in this setting may be to create a parity measure adjusted for demographics that incorporates a tolerance for verified differences in prevalence across protected groups.[46]

Equal opportunity and equalized odds metrics (based on TPR and FPR) are desirable criteria to satisfy because they represent equal performance accuracy of ML models across protected groups. However, these metrics are limited. Equal opportunity considers only TPR and fails to encapsulate other measures of diagnostic error or value, such as the positive predictive value of a model. The appropriate metric to optimize in a given context depends on the intended use case.[47] Metrics considered in this study may help identify important normative questions about decision making, as well as trade-offs and tensions between different potential interpretations of fairness.

Limitations

Our study has several limitations. The study is limited to 3 ML models (LR, SVM, and RNN models) trained to perform the specific task of predicting a future disease state given historical information and the disease state of individuals. It is not possible to extrapolate these results to fairness for other models or prediction tasks. Additionally, the study found unfairness in AD progression prediction, but it did not identify the source of unfairness in this context or how to mitigate it. Unfairness may arise due to features of the data or algorithms, and our investigation did not distinguish between these sources. Potential data biases include insufficient sample size in some groups and differential misclassification of disease stage and informative missingness.[35] Algorithmic bias arises when the bias is not present in the input data but is added purely by the algorithm.[48] It is generated by choices in the algorithmic design, including choice of predictor variables, optimization function, regularization, and loss function. Choices for each of these aspects of the algorithm may potentially bias the outcome of the algorithms.[17] Additionally, self-reported demographic and health information

Downloaded from jamanetwork.com by guest on 01/15/2024

has the potential to facilitate observational studies of AD but may introduce bias in results when groups in the cohort have different reporting approaches.[49-51]

Conclusions

In this prognostic study, 3 evaluated models performed well in aggregate but failed to satisfy metrics of fairness with respect to some of the protected attributes we investigated. Investigations of equal opportunity, equalized odds, and demographic parity found that models exhibited little unfairness with respect to sex but had notable deficits in fairness across race and ethnicity groups. This study highlights the potential for unfairness in ML-based AD prediction modeling and the importance of devoting attention to mitigating bias and advancing health equity. Future work will investigate mechanisms by which a model's design, data, and deployment may lead to disparities in AD. Developing a fairness-constrained model may be one avenue to address fairness challenges found in this study.

ARTICLE INFORMATION

Accepted for Publication: September 27, 2023.

Published: November 7, 2023. doi:10.1001/jamanetworkopen.2023.42203

Open Access: This is an open access article distributed under the terms of the CC-BY-NC-ND License. © 2023 Yuan C et al. *JAMA Network Open*.

Corresponding Author: Rebecca A. Hubbard, PhD (rhubb@pennmedicine.upenn.edu) and Kristin A. Linn, PhD (klinn@pennmedicine.upenn.edu), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Philadelphia, PA 19146.

Author Affiliations: Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Yuan, Linn, Hubbard); Penn Statistics in Imaging and Visualization Endeavor, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Yuan, Linn).

Author Contributions: Dr Yuan had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Linn and Hubbard contributed equally.

Concept and design: All authors.

Acquisition, analysis, or interpretation of data: Yuan, Linn.

Drafting of the manuscript: Yuan.

Critical review of the manuscript for important intellectual content: Linn, Hubbard.

Statistical analysis: All authors.

Obtained funding: Linn, Hubbard.

Supervision: Linn, Hubbard.

Conflict of Interest Disclosures: Dr Linn reported receiving personal fees from *JAMA Network Open* for providing statistical review outside the submitted work. Dr Hubbard reported receiving grants from Merck outside the submitted work. No other disclosures were reported.

Funding/Support: Research reported in this publication was supported by grant R21AG075574 from the National Institutes of Health (NIH). Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) by grants U01 AG024904 from the NIH and W81XWH-12-2-0012 from the Department of Defense. A complete listing of ADNI investigators can be found at the ADNI website. ADNI is funded by the National Institute on Aging and National Institute of Biomedical Imaging and Bioengineering and through generous contributions from AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc; Cogstate; Eisai Inc; Elan Pharmaceuticals, Inc; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research and Development, LLC; Johnson & Johnson Pharmaceutical Research and Development LLC; Lumosity; Lundbeck; Merck and Co, Inc; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and

Downloaded from jamanetwork.com by guest on 01/15/2024

Transition Therapeutics. The Canadian Institutes of Health Research provides funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This work also uses The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge data sets constructed by the European Progression Of Neurological Disease (EuroPOND) consortium, funded by grant 666992 from the European Union Horizon 2020 research and innovation program.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. *EBioMedicine*. 2022;84:104250. doi:10.1016/j.ebiom.2022.104250

2. Westman E, Muehlboeck JS, Simmons A. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *Neuroimage*. 2012;62(1):229-238. doi:10.1016/j.neuroimage.2012.04.056

3. Spasov S, Passamonti L, Duggento A, Liò P, Toschi N; Alzheimer's Disease Neuroimaging Initiative. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage*. 2019;189:276-287. doi:10.1016/j.neuroimage.2019.01.031

4. Huang L, Jin Y, Gao Y, Thung KH, Shen D; Alzheimer's Disease Neuroimaging Initiative. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiol Aging*. 2016;46:180-191. doi:10.1016/j.neurobiolaging.2016.07.005

5. Dickerson BC, Wolk DA; Alzheimer's Disease Neuroimaging Initiative. Biomarker-based prediction of progression in MCI: Comparison of AD signature and hippocampal volume with spinal fluid amyloid-β and tau. *Front Aging Neurosci*. 2013;5(OCT):55. doi:10.3389/fnagi.2013.00055

6. Franzmeier N, Koutsouleris N, Benzinger T, et al; Alzheimer's disease neuroimaging initiative (ADNI); Dominantly Inherited Alzheimer Network (DIAN). Predicting sporadic Alzheimer's disease progression via inherited Alzheimer's disease-informed machine-learning. *Alzheimers Dement*. 2020;16(3):501-511. doi:10.1002/alz.12032

7. Huan T, Tran T, Zheng J, et al. Metabolomics analyses of saliva detect novel biomarkers of Alzheimer's disease. *J Alzheimers Dis*. 2018;65(4):1401-1416. doi:10.3233/JAD-180711

8. de Leeuw FA, Peeters CFW, Kester MI, et al. Blood-based metabolic signatures in Alzheimer's disease. *Alzheimers Dement (Amst)*. 2017;8(1):196-207. doi:10.1016/j.dadm.2017.07.006

9. Zhang D, Shen D; Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage*. 2012;59(2):895-907. doi:10.1016/j.neuroimage.2011.09.069

10. Challis E, Hurley P, Serra L, Bozzali M, Oliver S, Cercignani M. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *Neuroimage*. 2015;112:232-243. doi:10.1016/j.neuroimage.2015.02.037

11. Lamb J, Crawford ED, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313(5795):1929-1935. doi:10.1126/science.1132939

12. Zissimopoulos JM, Barthold D, Brinton RD, Joyce G. Sex and race differences in the association between statin use and the incidence of Alzheimer disease. *JAMA Neurol*. 2017;74(2):225-232. doi:10.1001/jamaneurol.2016.3783

13. Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: evaluating claims and practices. In: *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2020:469-481. doi:10.1145/3351095.3372828

14. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med*. 2020;26(1):16-17. doi:10.1038/s41591-019-0649-2

15. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348

Downloaded from jamanetwork.com by guest on 01/15/2024

16. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature*. 2018;559(7714):324-326. doi:10.1038/d41586-018-05707-8

17. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1-35. doi:10.1145/3457607

18. Chen J, Kallus N, Mao X, Svacha G, Udell M. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In: *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery*; 2019:339-348. doi:10.1145/3287560.3287594

19. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. *Entropy (Basel)*. 2021;23(9):1165. doi:10.3390/e23091165

20. Shachar C, Gerke S. Prevention of bias and discrimination in clinical practice algorithms. *JAMA*. 2023;329(4):283-284. doi:10.1001/jama.2022.23867

21. Dworkin JD, Linn KA, Teich EG, Zurn P, Shinohara RT, Bassett DS. The extent and drivers of gender imbalance in neuroscience reference lists. *Nat Neurosci*. 2020;23(8):918-926. doi:10.1038/s41593-020-0658-y

22. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0

23. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*. 2022;6(12):1346-1352. doi:10.1038/s41551-022-00914-1

24. Sahin D, Jessen F. Kambeitz algorithmic fairness in biomarker-based machine learning models to predict Alzheimer's dementia in individuals with mild cognitive impairment. *Alzheimers Dement*. 2022;18:e062125. doi:10.1002/alz.062125

25. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

26. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176-2182. doi:10.1038/s41591-021-01595-0

27. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*. 2020;117(23):12592-12594. doi:10.1073/pnas.1919012117

28. El-Sappagh S, Alonso-Moral JM, Abuhmed T, et al. Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges. *Artif Intell Rev*. 2023;56:11149-11296. doi:10.1007/s10462-023-10415-5

29. Petti U, Nyrup R, Skopek JM, Korhonen A. Ethical considerations in the early detection of Alzheimer's disease using speech and AI. In: *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2023:1062-1075. doi:10.1145/3593013.3594063

30. El-Sappagh S, Ali F, Abuhmed T, Singh J, Alonso JM. Automatic detection of Alzheimer's disease progression: an efficient information fusion approach with heterogeneous ensemble classifiers. *Neurocomputing (Amst)*. 2022;512:203-224. doi:10.1016/j.neucom.2022.09.009

31. Jack CR Jr, Bernstein MA, Fox NC, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J Magn Reson Imaging*. 2008;27(4):685-691. doi:10.1002/jmri.21049

32. Marinescu RV, Oxtoby NP, Young AL, et al. TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease. *arXiv*. Preprint posted online May 10, 2018. doi:10.48550/arXiv.1805.03909

33. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383(9):874-882. doi:10.1056/NEJMms2004740

34. Verma S, Rubin J. Fairness definitions explained. In: *FairWare '18: Proceedings of the International Workshop on Software Fairness*. Association for Computing Machinery; 2018;1-7. doi:10.1145/3194770.3194776

35. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872. doi:10.7326/M18-1990

36. Nguyen M, He T, An L, Alexander DC, Feng J, Yeo BTT; Alzheimer's Disease Neuroimaging Initiative. Predicting Alzheimer's disease progression using deep recurrent neural networks. *Neuroimage*. 2020;222:117203. doi:10.1016/j.neuroimage.2020.117203

37. Handels RLH, Vos SJB, Kramberger MG, et al. Predicting progression to dementia in persons with mild cognitive impairment using cerebrospinal fluid markers. *Alzheimers Dement*. 2017;13(8):903-912. doi:10.1016/j.jalz.2016.12.015

Downloaded from jamanetwork.com by guest on 01/15/2024

38. Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neurosci*. 2019;11:220. doi:10.3389/fnagi.2019.00220

39. Albright J. Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. *Alzheimers Dement (N Y)*. 2019;5(1):483-491. doi:10.1016/j.trci.2019.07.001

40. Mehdipour Ghazi M, Nielsen M, Pai A, et al; Alzheimer's Disease Neuroimaging Initiative. Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. *Med Image Anal*. 2019;53:39-46. doi:10.1016/j.media.2019.01.004

41. Casanova R, Barnard RT, Gaussoin SA, et al; WHIMS-MRI Study Group and the Alzheimer's disease Neuroimaging Initiative. Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases. *Neuroimage*. 2018;183:401-411. doi:10.1016/j.neuroimage.2018.08.040

42. Bayle P, Bayle A, Janson L, Mackey L. Cross-validation confidence intervals for test error. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems. Vol 33*. Curran Associates, Inc; 2020:16339-16350. Accessed October 3, 2023. https://proceedings.neurips.cc/paper_files/paper/2020/file/bce9abf229ffd7e570818476ee5d7dde-Paper.pdf

43. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front Artif Intell*. 2021;3:561802. doi:10.3389/frai.2020.561802

44. Lennon JC, Aita SL, Bene VAD, et al. Black and White individuals differ in dementia prevalence, risk factors, and symptomatic presentation. *Alzheimers Dement*. 2022;18(8):1461-1471. doi:10.1002/alz.12509

45. Power MC, Bennett EE, Turner RW, et al. Trends in relative incidence and prevalence of dementia across non-Hispanic Black and White individuals in the United States, 2000-2016. *JAMA Neurol*. 2021;78(3):275-284. doi:10.1001/jamaneurol.2020.4471

46. 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

47. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*. 2021;4:123-144. doi:10.1146/annurev-biodatasci-092820-114757

48. Danks D, London AJ. Algorithmic bias in autonomous systems. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization; 2017:4691-4697. doi:10.24963/ijcai.2017/654

49. Ashford MT, Neuhaus J, Jin C, et al. Predicting amyloid status using self-report information from an online research and recruitment registry: The Brain Health Registry. *Alzheimers Dement (Amst)*. 2020;12(1):e12102. doi:10.1002/dad2.12102

50. Fowler ME, Triebel KL, Cutter GR, Schneider LS, Kennedy RE; Alzheimer's Disease Neuroimaging Initiative. Progression of Alzheimer's disease by self-reported cancer history in the Alzheimer's Disease Neuroimaging Initiative. *J Alzheimers Dis*. 2020;76(2):691-701. doi:10.3233/JAD-200108

51. Kuhn E, Perrotin A, La Joie R, et al; Alzheimer's Disease Neuroimaging Initiative. Association of the informant-reported memory decline with cognitive and brain deterioration through the Alzheimer clinical continuum. *Neurology*. 2023;100(24):e2454-e2465. doi:10.1212/WNL.0000000000207338

SUPPLEMENT 1.

eMethods.
eFigure 1. Overview of Model Pipeline
eFigure 2. Absolute Differences in True Positive Rates Across Groups Defined by 3 Protected Attributes
eFigure 3. True Positive Rates of Mild Cognitive Impairment by Protected Attribute
eFigure 4. False Positive Rates of Mild Cognitive Impairment by Protected Attribute
eFigure 5. Predicted Probability of Progression by Protected Attribute
eFigure 6. Differences of Predicted Progression Probabilities by Protected Attribute
eTable 1. Summary Statistics for Protected Attributes and Predictor Variables by Cognitive Functioning Trajectory for Excluded Trajectories
eTable 2. Mean Prediction Performance Across 10 Test Sets

SUPPLEMENT 2.

Data Sharing Statement

Downloaded from jamanetwork.com by guest on 01/15/2024