Research Article

Predicting cognitive decline from neuropsychiatric symptoms and Alzheimer's disease biomarkers: A machine learning approach to a population-based data

Journal of Alzheimer's Disease
2025, Vol. 103(3) 833–843

DOI: 10.1177/13872877241306654
journals.sagepub.com/home/alz

Jay Shah[1,2], Janina Krell-Roesch[3,4], Erica Forzani[5], David S Knopman[6], Cliff R Jack[7], Ronald C Petersen[3,6], Yiming Che[1,2], Teresa Wu[1,2] and Yonas E Geda[8]

Abstract

Background: The aim of this study was to examine the potential added value of including neuropsychiatric symptoms (NPS) in machine learning (ML) models, along with demographic features and Alzheimer's disease (AD) biomarkers, to predict decline or non-decline in global and domain-specific cognitive scores among community-dwelling older adults.

Objective: To evaluate the impact of adding NPS to AD biomarkers on ML model accuracy in predicting cognitive decline among older adults.

Methods: The study was conducted in the setting of the Mayo Clinic Study of Aging, including participants aged ≥ 50 years with information on demographics (i.e., age, sex, education), NPS (i.e., Neuropsychiatric Inventory Questionnaire; Beck Depression and Anxiety Inventories), at least one AD biomarker (i.e., plasma-, neuroimaging- and/or cerebrospinal fluid [CSF]-derived), and at least 2 repeated neuropsychological assessments. We trained and tested ML models using a stepwise feature addition approach to predict decline versus non-decline in global and domain-specific (i.e., memory, language, visuospatial, and attention/executive function) cognitive scores.

Results: ML models had better performance when NPS were included along with a) neuroimaging biomarkers for predicting decline in global cognition, as well as language and visuospatial skills; b) plasma-derived biomarkers for predicting decline in visuospatial skills; and c) CSF-derived biomarkers for predicting decline in attention/executive function, language, and memory.

Conclusions: NPS, added to ML models including demographic and AD biomarker data, improves prediction of downward trajectories in global and domain-specific cognitive scores among community-dwelling older adults, albeit effect sizes are small. These preliminary findings need to be confirmed by future cohort studies.

Keywords

Alzheimer's disease biomarkers, cognitive trajectory, machine learning, neuropsychiatric symptoms

Received: 26 June 2024; accepted: 27 October 2024

[1]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA
[2]ASU-Mayo Center for Innovative Imaging, Tempe, AZ, USA
[3]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA
[4]Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
[5]Biodesign Institute, Arizona State University, Tempe, AZ, USA
[6]Department of Neurology, Mayo Clinic, Rochester, MN, USA
[7]Department of Radiology, Mayo Clinic, Rochester, MN, USA
[8]Department of Neurology and the Franke Barrow Global Neuroscience Education Center, Barrow Neurological Institute, Phoenix, AZ, USA

Part of this work was presented at the American Academy of Neurology (AAN) annual meeting 2023 in Boston, MA, USA.

Corresponding author:
Teresa Wu, School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA.
Email: Teresa.Wu@asu.edu

Introduction

Neuropsychiatric symptoms (NPS) such as apathy or depression are very common in older adults, with prevalence rates ranging between 25% in cognitively unimpaired individuals to about 50% in those with mild cognitive impairment (MCI).[1,2] In persons with Alzheimer's disease (AD), almost all patients develop NPS at some point during the course of the illness.[3] NPS are a well-established risk factor for cognitive impairment and are associated with an increased risk of incident MCI,[4,5] dementia,[6–10] and decline in cognitive trajectories.[11,12]

AD biomarkers derived from neuroimaging, cerebrospinal fluid (CSF), and, more recently, blood provides the opportunity for in vivo investigation of NPS in biologically-defined AD.[13] To this end, a growing body of research has shown that NPS coupled with AD biomarker abnormality may be associated with a higher risk of cognitive impairment or decline than either alone.[14–16]

To date, most research on the associations between NPS and AD biomarkers in predicting cognitive impairment or decline has been conducted using traditional statistical approaches, and there is a lack of studies that leverage machine learning (ML) based approaches to study these associations. This is important as ML models may have added value in capturing complex and non-linear patterns, particularly in larger datasets. ML is increasingly being used in brain aging research and may aid in making early diagnosis of AD and increase the specificity and sensitivity of the diagnosis.[17,18] Studies have also shown that neuroimaging data can be used to study changes in the brain suggestive of AD using different ML techniques with acceptable accuracy, offering the potential for individualized diagnosis.[19,20] However, in a recently published literature review by our group,[21] we found only one study using Alzheimer's Disease Neuroimaging Initiative (ADNI) data that utilized ML to combine NPS and AD biomarker information in predicting cognitive impairment, i.e., classification of normal cognition and MCI/AD.[22] This study, however, focused only on using neuroimaging biomarkers of AD and baseline information from NPI-Q to determine mild behavioral impairment (MBI) and MBI domains.

Thus, the aim of this study was to develop ML models to predict decline or non-decline in global and domain-specific cognitive scores from AD biomarkers and NPS in a large sample of community-dwelling older adults. Specifically, we examined whether ML models including demographic features (i.e., sex, education, and age) and (1) plasma-derived biomarkers and NPS; (2) neuroimaging-derived biomarkers and NPS; and (3) CSF-derived biomarkers and NPS predicted decline or non-decline in global and domain-specific (i.e., i.e., memory, language, visuospatial, and attention/ executive function) cognitive scores; and whether model performance was better with or without NPS features included.

Methods

Study setting and participants

Our study was conducted in the setting of the population-based Mayo Clinic Study of Aging in Olmsted County, MN, USA. The reader is referred to Roberts et al.[23] for details on the study methodology and procedures. Briefly, all participants undergo a face-to-face evaluation, including a neurological examination and administration of the Short Test of Mental Status,[24] a study coordinator visit to assess sociodemographic data, NPS, and activities of daily living, and neuropsychological testing supervised by a psychometrist to assess performance in four cognitive domains, i.e., memory, language, visuospatial skills, and attention/ executive function. An expert consensus panel of physicians, study coordinators, and neuropsychologists then reviews the results for each participant and determines whether a participant is cognitively unimpaired or has cognitive impairment (i.e., MCI or dementia). This classification is done based on published normative data developed on a different sample in this community. For MCI, the revised Mayo Clinic criteria are used.[25] Participants with MCI have a Clinical Dementia Rating (CDR) score of 0 or 0.5; however, the final diagnosis of MCI is based on all available data. For the analyses reported in this manuscript, we included Mayo Clinic Study of Aging participants aged 50 years and older (refer to Table 1 for cognitive status summary), on whom information about NPS and at least one AD biomarker, and repeated assessments (at least 2) of global and domain-specific cognitive function were available. All samples had PiB-PET and FDG-PET, a portion of those samples also had plasma biomarkers, and a portion of the sample also had CSF biomarkers. The study was approved by the Mayo Clinic and Olmsted Medical Center institutional review boards, and written informed consent was obtained from every participant. In the case of participants with cognitive impairment sufficient to interfere with capacity, assent was obtained from a legally authorized representative.

Neuropsychological assessment

We assessed performance in four cognitive domains using standardized and validated neuropsychological tests: (1) memory: Auditory Verbal Learning Test - delayed recall,[26] and Wechsler Memory Scale-Revised Logical Memory II - delayed recall and Visual Reproduction II Tests - delayed recall;[27] (2) language: Boston Naming Test,[28] and category fluency;[29]; (3) visuospatial skills: Wechsler Adult Intelligence Scale-Revised Block Design and Picture Completion Tests;[30] (4) and attention/executive function: Trail-Making Test Part B,[31] and Wechsler Adult Intelligence Scale-Revised Digit Symbol Substitution Test.[30] Using these individual tests, global and domain-

Table 1. Description of participant demographics included in this study for the three sets of ML experiments.

Experiment	Total Sample	Sex (Male /Female)	Education (Mean ± SD y)	Age (Mean ± SD y)	Baseline versus follow-up time difference (Mean ± SD, median y)	Baseline versus follow-up change in cognitive outcome (Mean ± SD, median)	Cognitive Status	*APOE* ϵ4 (carrier/ non-carrier)
(1) Using demographics, plasma-derived biomarkers and NPS	1254	531/723	14.9 ± 2.6	70.9 ± 9.4	6.0 ± 2.9, 6.5	−0.11 ± 0.72, −0.02	1126 CU, 104 MCI, 22 dementia, 2 missing	332/915, 7 missing
(2) Using demographics, neuroimaging-derived biomarkers and NPS	1345	563/782	14.9 ± 2.6	71.0 ± 9.5	6.0 ± 3.0, 6.5	−0.11 ± 0.72, −0.04	1203 CU, 117 MCI, 23 dementia, 2 missing	356/980, 9 missing
(3) Using demographics, CSF-derived biomarkers and NPS	299	120/179	14.8 ± 2.5	68.8 ± 9.2	6.7 ± 2.8, 7.5	−0.06 ± 0.76, 0.07	267 CU, 25 MCI, 6 dementia, 1 missing	91/205, 3 missing

CU: cognitively unimpaired; MCI: mild cognitive impairment. Samples with missing cognitive status had validity issues with the diagnosis. This table summarizes the maximum number of samples for a given experiment with no missing entries. However, the sample count does vary for different outcomes (refer to Tables 3, 4, and 5).

specific cognitive z-scores were calculated as detailed by Pink et al.[15] An expert consensus panel of physicians, study coordinators, and neuropsychologists reviewed the results for each participant and determined whether a participant was cognitively unimpaired or had cognitive impairment (i.e., MCI or dementia). In this manuscript, we used the change between the baseline and the last visit in global cognition z-score as well as in domain-specific cognitive z-scores (i.e., memory, language, visuospatial skills, attention/ executive function) to determine decline versus non-decline as the outcomes of interest in our analyses.

AD biomarker ascertainment

The Mayo Clinic Study of Aging team has published its methodology on plasma[32,33] and CSF-derived[34] biomarker assessment. Briefly, plasma-derived $A\beta_{42}/A\beta_{40}$ ratio, glial fibrillary acidic protein (GFAP), and neurofilament light chain (NfL) were quantified via the Simoa® Neurology 4-Plex E Advantage kit (N4PE, item #103,670), and for phospho-tau 181 (p-tau181), the Simoa® p-tau181 Advantage V2 kit (item #103,714) was utilized. All samples were measured on a Quanterix HD-X analyzer (Quanterix, Lexington, MA, USA). CSF amyloid-beta1-42 ($A\beta_{42}$), total Tau (t-tau) and hyperphosphorylated Tau-181 (p-tau181) were analyzed using Elecsys® β-Amyloid (1-42) CSF, Elecsys® Total-Tau CSF, and Elecsys® Phospho-Tau (181P) CSF electrochemiluminescence immunoassays (Roche Diagnostics). Furthermore, the Mayo Clinic Study of Aging team has published its methodology including cut points, on FDG-PET (<=1.47),[35–37] and amyloid (PiB PET) imaging (>=1.48).[35,38–40] In this manuscript, we used four plasma-derived (i.e., $A\beta_{42}/A\beta_{40}$ ratio, GFAP, NfL, and p-tau181), three CSF-derived (i.e., $A\beta_{42}$, t-tau, and p-tau181), and two neuroimaging-derived biomarkers (i.e., PiB-PET and FDG-PET) for ML-based prediction of cognitive trajectories.

Neuropsychiatric symptoms assessment

NPS were assessed using the Neuropsychiatric Inventory Questionnaire (NPI-Q[41]), which was administered as a structured interview to an informant by a study coordinator and assessed the presence or absence of 12 symptoms (i.e., depression, anxiety, apathy, agitation, delusions, hallucinations, euphoria, disinhibition, irritability, aberrant motor behavior, sleep/ nighttime disturbance behavior, and eating/appetite). In addition, we assessed self-reported depression and anxiety using the Beck Depression Inventory II (BDI-II[42]) and Beck Anxiety Inventory (BAI[43]). The BDI-II measures common symptoms of depression, such as feelings of guilt or loss of interest, over the preceding 2 weeks. The BAI measures common anxiety symptoms, such as nervousness or fear of losing control, over the preceding week. Both inventories are

Table 2. Number of features input to ML models for each predictor in each experiment in stepwise feature addition approach.

Experiment	Step	Predictors (number of features)	Cognitive outcomes
1	1	demographics (3)	(a) attention/ executive function
	2	demographics (3), plasma-derived biomarkers (4)	(b) visuospatial
	3	demographics (3), plasma-derived biomarkers (4), neuropsychiatric symptoms (14)	(c) language
2	1	demographics (3)	(d) memory
	2	demographics (3), neuroimaging-derived biomarkers (2)	(e) global
	3	demographics (3), neuroimaging-derived biomarkers (2), neuropsychiatric symptoms (14)	
3	1	demographics (3)	
	2	demographics (3), CSF-derived biomarkers (3)	
	3	demographics (3), CSF-derived biomarkers (3), neuropsychiatric symptoms (14)	

validated and consist of 21 items. The severity of each item is rated on a Likert-type scale ranging from 0 to 3, with the total score ranging from 0 to 63 and a higher score indicating higher symptom severity. In this manuscript, we used 14 measures of NPS (i.e., presence/ absence of 12 NPI-Q-derived NPS, BAI total score, and BDI-II total score) for ML-based prediction of cognitive trajectories. NPS measured by these scales consider that behavioral traits such as anxiety and impulsivity are indeed inherently subjective.

Machine learning (Ml) models

We trained and tested ML models using a stepwise feature addition approach to predict trajectories in one global and four domain-specific (i.e., memory, language, visuospatial, and attention/ executive function) cognitive outcomes. Table 2 provides an overview of the experimental design and predictors included in the prediction of cognitive outcomes, and Table 1 describes a brief summary of sample demographics for three sets of ML experiments. We discarded samples from data that had missing values for any of the measurements. The data included after cleaning missing value entries contained five different sets of measurements, namely: (1) demographics - sex, education, and age; (2) plasma-derived biomarkers - $A\beta_{42}/A\beta_{40}$ ratio, GFAP, NfL, and p-tau181; (3) neuroimaging-derived biomarkers - PiB-PET and FDG-PET; (4) CSF-derived biomarkers - $A\beta_{42}$, t-tau, and p-tau181; (5) NPS – BDI-II total score (continuous variable; possible range: 0–63 with higher score indicating higher depressive symptoms), BAI total score (continuous variable; possible range: 0–63 with higher score indicating higher anxiety symptoms), 12 NPI-Q features (categorical variables; presence or absence of the respective neuropsychiatric symptom). We used each participant's first and last recorded visit data to determine if there was a decline or non-decline in cognitive outcome variables.

To study the role of NPS in predicting decline or non-decline in cognitive outcome variables, we conducted three sets of experiments and used a stepwise approach to build predictive models as illustrated in Figure 1: (1) including demographic features, then adding plasma-derived biomarkers, then adding NPS; (2) including demographic features, then adding neuroimaging-derived biomarkers, then adding NPS; and (3) including demographic features, then adding CSF-derived biomarkers, then adding NPS. All features were pre-processed by removing the mean and scaling to unit variance before inputting it into ML models.[44] We use 5-fold cross-validation with grid search "GridSearchCV" technique[45] for finding the optimal parameter configuration from a given set of parameters in a grid (see Supplemental Table 1). Data is split in a ratio of 90:10 for training and held-out testing sets, respectively, while maintaining the same distribution of samples with decline and non-decline in cognitive outcomes. GridSearchCV performs a 5-fold cross-validation on the training data for hyperparameter optimization. For a given ML model, training fold, and different hyperparameters - GridSearchCV finds the best parameters using AUC as a validation metric. Among 10 different ML models (see paragraph below), using the best hyperparameter configuration, a single model is selected that performs best among five folds of training data. This model is re-trained on entire training data, and predictions are made on the held-out test set. For each cognitive outcome variable, we only chose the best-performing ML model using demographic features and compared its performance when NPS and other biomarkers were added.

Different ML algorithms learn differently from the underlying data, which is known to introduce strengths or weaknesses.[46] To avoid algorithm specific limitations, we included 10 different ML models with best parameters obtained from GridSearchCV for each experiment, namely (1) Logistic Regression—a generalized linear ML model where estimated probability response is a linear function of predictor parameters; (2) K-nearest neighbors - non-parametric supervised learning algorithm that finds the k most similar training examples to a new data point and then uses the labels of those training examples to predict the label of the new data point; (3) Support Vector Machine - used to classify data by finding the optimal

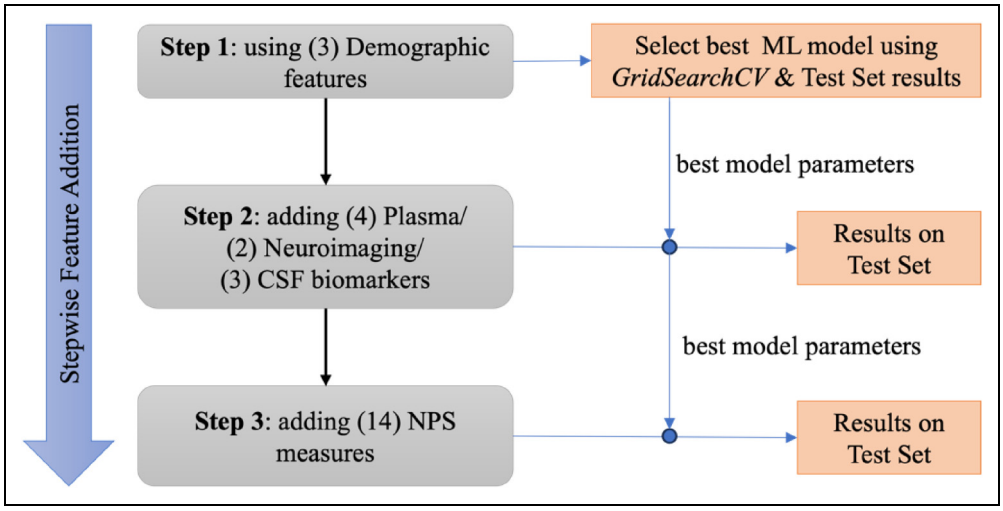

Figure 1. Building different sets of ML models using a stepwise feature addition approach.

decision boundary that separates different classes and maximizes the margin between support vectors; (4) Gaussian Process - Bayesian method that makes predictions with uncertainty; (5) Decision Trees - predicts outcomes by learning and mapping decisions in a branching structure; (6) Random Forests - ensembling technique that combines outputs of multiple decision trees to make predictions; (7) Multi-layer Perceptron - a simple feedforward neural network that can learn non-linear associations between input data and target variables; (8) Naive Bayes - a probabilistic classifier that predicts outcomes based on probability of occurrence of features; (9) Quadratic Discriminant Analysis - assumes and learns a gaussian distribution for each class in target/outcome variables space; and (10) eXtreme Gradient Boosting (XGBoost) - uses gradient boosted combination of decision-trees to make predictions. The performance of each model was evaluated using the area under the ROC curve (AUC), as well as brier score, accuracy, sensitivity, specificity, and precision. Accuracy was defined as (TP + TN)/N, where TP is the number of true positives and TN for true negatives from total N samples. Sensitivity and Specificity measure the ability of a classifier to identify positive and negative instances, here, subjects with cognitive decline and those without, respectively. Sensitivity was defined as TP/(TP + FN) and specificity as TN/(TN + FP), where FN and FP are falsely classified as negative and positive, respectively.

We added brier score to evaluate the potential added value of including NPS in ML models along with demographic, CSF, plasma and neuroimaging biomarkers. Net reclassification index (NRI) and DeLong's test are two different statistical methods to compare risk prediction models. However, NRI is known to produce overly optimistic results when risk models do not adequately fit the test data,[47,48] which can result in seemingly improved prediction performance based on positive NRI values when evaluated on the test data—even if the models do not genuinely enhance prediction accuracy. Whereas, in DeLong's test, the inherent assumption of a Gaussian distribution for the empirical AUC differences between the full and reduced models can be overly restrictive, leading to conservative results, particularly in cases involving nested models, as in our study.[49] Pepe et al. (2014)[47] suggested that the AUC difference (ΔAUC) between the full and reduced models and the Brier Score difference (ΔBrier) demonstrate more consistent performance in reflecting true prediction improvement. Additionally, we employed the Wald test to demonstrate the significance of the added NPS features.

Results

Predictive performance of global and domain-specific cognitive outcomes

Table 3 shows the predictive performance of decline or non-decline in global and domain-specific cognitive scores using different ML models, including demographic features, plasma-derived biomarkers, and NPS. Besides memory and visuospatial, adding plasma biomarkers and NPS improved the ML model's overall performance (measured using AUC) in differentiating between increasing or decreasing cognitive trajectories. For language, even though AUC did not improve, specificity improved significantly (0.91 versus 0.12), demonstrating the effectiveness of NPS in reducing false positives in ML model predictions. For global cognition prediction, adding NPS to demographics and plasma biomarkers had no positive effect on model performance.

Similarly, adding NPS to demographic features and neuroimaging biomarkers improved ML models' classification performance for all outcomes of interest in our dataset, except for attention (please refer to Table 4). In particular,

Table 3. Performance of ML models in predicting decline (versus non-decline) in global and domain-specific cognitive outcomes using stepwise feature addition (plasma-derived biomarkers and NPS).

Cognitive Outcomes	Samples	Step	ML model	ACC	SPE	SEN	PRE	BRI	AUC
Attention/Executive function	1222 (374/848)	1	Logistic Regression	0.67	0.19	0.88	0.72	0.19	0.66
		2		0.71	0.16	0.94	0.72	0.15	**0.75**
		3		0.70	0.00	1.00	0.70	0.16	0.71
Visuospatial	1228 (571/657)	1	Logistic Regression	0.64	0.61	0.68	0.60	0.24	0.66
		2		0.59	0.56	0.63	0.55	0.22	**0.68**
		3		0.63	0.58	0.68	0.58	0.24	0.66
Language	1231 (463/768)	1	Naive Bayes	0.60	0.02	0.97	0.61	0.38	0.59
		2		0.62	0.58	0.64	0.71	0.23	**0.63**
		3		0.48	0.92	0.21	0.80	0.23	0.62
Memory	1254 (599/659)	1	SVM	0.68	0.57	0.78	0.68	0.21	**0.73**
		2		0.62	0.53	0.69	0.64	0.21	0.71
		3		0.63	0.57	0.68	0.65	0.21	0.71
Global	1202 (489/713)	1	Random Forest	0.69	0.58	0.78	0.71	0.21	0.73
		2		0.66	0.35	0.90	0.65	0.22	**0.76**
		3		0.64	0.38	0.84	0.64	0.22	0.74

Step 1: Using demographic features only; Step 2: adding plasma-derived biomarkers; Step 3: adding NPS. Bold values indicate the best performance. Samples: total count (increasing/decreasing cognitive outcome measures); ACC: accuracy; SPE: specificity; SEN: sensitivity; PRE: precision; BRI: brier score; AUC: area under curve; SVM: support vector machine; XGB: extreme gradient boosting.

Table 4. Performance of ML models in predicting decline (versus non-decline) in global and domain-specific cognitive outcomes using stepwise feature addition approach (neuroimaging-derived biomarkers and NPS).

Cognitive Outcomes	Samples	Step	ML model	ACC	SPE	SEN	PRE	BRI	AUC
Attention/Executive function	1312 (405/907)	1	Decision Tree	0.77	0.42	0.88	0.82	0.16	**0.72**
		2		0.75	0.27	0.91	0.79	0.18	0.64
		3		0.77	0.42	0.88	0.82	0.18	0.69
Visuospatial	1317 (612/705)	1	Random Forest	0.62	0.48	0.74	0.63	0.23	0.68
		2		0.62	0.47	0.75	0.63	0.22	0.70
		3		0.63	0.45	0.78	0.63	0.22	**0.72**
Language	1322 (503/819)	1	XGB	0.65	0.26	0.89	0.67	0.21	0.72
		2		0.66	0.28	0.89	0.67	0.20	0.71
		3		0.71	0.38	0.90	0.71	0.19	**0.75**
Memory	1345 (633/712)	1	QDA	0.70	0.58	0.78	0.73	0.30	0.71
		2		0.67	0.67	0.68	0.75	0.20	**0.74**
		3		0.57	0.91	0.34	0.84	0.21	0.73
Global	1290 (524/766)	1	XGB	0.64	0.43	0.79	0.67	0.22	0.67
		2		0.64	0.43	0.78	0.66	0.22	0.69
		3		0.66	0.45	0.80	0.68	0.21	**0.70**

Step 1: Using demographic features only; Step 2: adding neuroimaging-derived biomarkers; Step 3: adding NPS. Bold values indicate the best performance. Samples: total count (increasing/decreasing cognitive outcome measures); ACC: accuracy; SPE: specificity; SEN: sensitivity; PRE: precision; BRI: brier score; AUC: area under curve; QDA: quadratic discriminant analysis; XGB: extreme gradient boosting.

adding NPS considerably improved AUC for all cognitive outcomes except attention/executive function and memory. For memory, even though AUC did not improve, the specificity of the model had significant improvements (0.91 versus 0.67) with the added NPS in the ML model. Performance improvement might be domain specific, i.e., improved prediction in global cognition might be driven by improvements in visuospatial and language outcomes.

Furthermore, as shown in Table 5, there are improvements in AUC performance when adding NPS to demographic features and CSF-derived biomarkers for all cognitive outcomes, except for visuospatial skills outcomes. For language and global cognitive outcomes, including NPS significantly improved AUC scores.

For a visual display of data, please refer to the AUC plots in Supplemental Figures 1–3.

Discussion

We observed that adding NPS to the demographic and AD biomarkers ML model improved the prediction of decline in

Table 5. Performance of ML models in predicting decline (versus non-decline) in global and domain-specific cognitive outcomes using stepwise feature addition approach (CSF-derived biomarkers and NPS).

Cognitive Outcomes	Samples	Step	ML model	ACC	SPE	SEN	PRE	BRI	AUC
Attention/Executive function	296 (99/196)	1	Decision Tree	0.73	0.17	0.88	0.81	0.18	0.52
		2		0.60	0.50	0.63	0.83	0.17	0.56
		3		0.60	0.50	0.63	0.83	0.17	**0.56**
Visuospatial	294 (151/143)	1	XGB	0.77	0.89	0.58	0.78	0.21	0.72
		2		0.77	0.83	0.67	0.73	0.16	**0.81**
		3		0.73	0.83	0.58	0.70	0.16	0.80
Language	296 (125/171)	1	Neural Net	0.60	0.50	0.67	0.67	0.25	0.55
		2		0.57	0.42	0.67	0.63	0.25	0.56
		3		0.63	0.67	0.61	0.73	0.23	**0.65**
Memory	299 (164/135)	1	Neural Net	0.77	0.88	0.62	0.80	0.19	0.85
		2		0.73	0.76	0.69	0.69	0.19	0.79
		3		0.80	0.76	0.85	0.73	0.17	**0.85**
Global	296 (108/188)	1	Naive Bayes	0.70	0.67	0.70	0.95	0.24	**0.72**
		2		0.67	0.33	0.70	0.90	0.38	0.63
		3		0.67	0.67	0.67	0.95	0.25	0.69

Step 1: Using demographic features only; Step 2: adding CSF-derived biomarkers; Step 3: adding NPS. Bold values indicate the best performance. Samples: total count (increasing/decreasing cognitive outcome measures); ACC: accuracy; SPE: specificity; SEN: sensitivity; PRE: precision; BRI: brier score; AUC: area under curve; XGB: extreme gradient boosting.

global and domain-specific cognitive scores, albeit effect sizes are small and domain specific. ML models performed better when NPS was included along with AD neuroimaging biomarkers to predict a decline in global cognition, and language and visuospatial domains. These results are consistent with findings from Gill et al.[22] who report added advantage of ML with NPS and imaging biomarkers for future diagnostic predictions (normal controls versus MCI/AD). Mallo et al.[50] also observed that ML algorithms can predict the risk of conversion from MCI to incident dementia (AUC: 0.85 ± 0.16) using socio-demographic data and NPS proxies. Similarly, in a study among patients with Parkinson's disease,[51] the best performing model to predict longitudinal cognitive decline included baseline cognition, CSF-derived biomarkers, and depression and anxiety.

We further observed that the predictive performance of ML models, including plasma-biomarkers, improved when NPS were included for the outcome of visuospatial skills and that the predictive performance of ML models, including CSF-derived biomarkers improved when NPS were included for the outcomes of attention/ executive function, language, and memory. To the best of our knowledge, there is no other published research that used ML to predict downward cognitive trajectory when NPS is added to demographic features and plasma- or CSF-derived AD biomarkers. Thus, it is not possible to compare our findings to other studies.

We also observed that for the outcome of language, adding NPS to ML models, including plasma-derived biomarkers, improved model specificity, i.e., the model's ability to detect true negatives. We also observed similar improvements in the specificity for ML model including neuroimaging biomarkers for the outcome of memory when adding NPS. To evaluate the statistical significance of AUC improvements, we show the distribution of empirical AUC differences between the models: (1) using demographic features only, (2) using demographics and CSF (3) using demographics, CSF and NPS features using the Neural Net model to predict change in language outcome in Supplemental Figure 4. It is evident that the distribution is not Gaussian, which makes DeLong's test unsuitable.[49] Supplemental Tables 2–4 show the p-values of NPS features obtained from the Wald test with ordinary least squares (OLS) model, highlighting that while not all NPS features contribute significantly, several of them are indeed non-negligible.

Of note, when only considering AUC, adding NPS to the models of CSF- and neuroimaging-derived biomarkers, mainly improved prediction of language as compared to other cognitive domains. This may be surprising given that the language domain is generally least sensitive to early AD pathology. In contrast, memory and attention/ executive function are rather sensitive markers of early cognitive change. Thus, future research utilizing ML models should particularly examine the potential added value of NPS in predicting changes in memory and attention/executive function.

Overall, our research provides preliminary findings that NPS added to ML models including demographic and AD biomarker data improves prediction of decline in global and domain-specific cognitive scores in a large sample of community-dwelling older adults. This finding may inform the field of the neuropsychiatry of AD that investigates multiple NPS such as apathy, agitation, depression etc. as exposure of interest and AD as outcome of interest. These preliminary findings could have many future directions: (1) ML modeling aspects may include exploring

model uncertainties and imputation for missing data, and (2) from a clinical aspect, more research on the relevant features for clinical interpretations may be warranted. One must also acknowledge the inherent complexity of ML-based algorithms in comparison to traditional statistical models (e.g., multi-variate regression) that are widely used in research on NPS in the context of brain aging and AD research. Hence, understanding how a model makes prediction, and interpretation on which predictors are most informative in distinguishing between participants who experience cognitive decline over time compared to those who do not, would be of value. To this end, Lundberg et al.[52] proposed a post-hoc model interpretability algorithm SHapley Additive exPlanations (SHAP), using additive attribution to calculate which features contribute most to a model's predictions. Using SHAP in the future, we plan to further interrogate the features contributing to cognitive decline, and particularly the additional role of NPS for a deeper clinical understanding.

The strength of this study is that we used a large dataset containing information on different AD biomarker modalities, NPS, and cognitive trajectories among community-dwelling adults aged 50 years and older who are enrolled in the Mayo Clinic Study of Aging. To the best of our knowledge, this is one of the first studies to apply an ML-based approach in a sample of community-dwelling older adults to predict downward cognitive trajectory using NPS and AD biomarkers derived from neuroimaging, CSF, and plasma.

Limitations of our study pertain to the fact that we used traditional ML methods rather than advanced ML algorithms such as deep neural networks that are known to outperform traditional ML methods and even humans, e.g., in specific image analysis tasks.[21] These networks perform automatic feature extraction as opposed to most ML methods. Supplementing imaging data with clinical information can significantly improve diagnostic performance with deep neural networks.[53] In this study, we only used two neuroimaging measures derived from FDG and PiB-PET scans which can limit performance of ML model. At the same time, lumping together FDG and PiB-PET scans may also be considered a limitation itself, since they measure different brain changes due to AD, i.e., FDG-PET assesses abnormalities in brain glucose metabolism, and PiB-PET assesses cortical amyloid deposition. Thus, future work should focus on developing a deep learning-based model leveraging raw neuroimaging data and NPS to improve prediction of trajectories in cognitive outcomes, and should preferably only combine neuroimaging modalities that measure the same pathophysiological changes related to AD. Furthermore, behavioral traits such as anxiety and impulsivity are inherently subjective and may vary across different populations thereby limiting the generalizability of our findings. However, we used validated scales such as NPI-Q, BDI-II and BAI that take such measurement challenges into account. Nevertheless, it must be noted that collecting NPS via self-reported questionnaire from the individual or a caregiver might be biased by social desirability. In addition, it is important to note that no particular ML model outperforms in each experiment. This is mainly due to different sets of assumptions made by each model about the training data distribution, resulting in inductive bias.[54] Therefore, there is a need for further exploration of different models that can make robust predictions across different experiments. Vascular and other medical comorbidities may have an impact on cognitive decline in our study sample. However, we did not include medical comorbidities as confounding variables in our current analyses. We plan to build on this study to examine the potential impact of vascular risk factors (ATN-V model) in similar models in the future. We also note few limitations pertaining to our study settings: (1) Only a small portion of the sample had CSF data available (refer to Table 5) which can limit ML models' predictive performance; (2) our study sample is relatively highly educated, wealthy, and less ethnically diverse - about 99% of our study participants are White. Although it has been shown that data from Olmsted County are generalizable to the population of Minnesota and the upper Midwest,[55] our findings may not apply to the overall U.S. population; (3) the plasma biomarker field is rapidly advancing and utilizing ML models in analyzing plasma biomarker data measured on new assays/ platforms will be needed; (4) only categorical outcomes (decline/non-decline in cognitive trajectories) were used here, and we did not examine whether declines were clinically meaningful. Thus, future ML models should also explore the possibility of incorporating continuous cognitive test outcomes to provide more information about the quantitative nature and degree of cognitive decline over time.

In conclusion, the results of our study provide preliminary findings of a potential added value of considering NPS information along with AD biomarkers and demographic features in predicting decline in cognitive trajectories among community-dwelling adults aged 50 years and older by using a ML-based approach, albeit effect sizes are small. More research is needed to confirm these preliminary observations, which may have implications for clinical practice and further support the importance of a thorough neuropsychiatric assessment, and potential treatment of NPS, among older adults at risk for cognitive impairment and/ or with known AD biomarker abnormality.

Acknowledgments

The authors have no acknowledgments to report.

ORCID iDs

Jay Shah https://orcid.org/0000-0002-3617-2395
Teresa Wu https://orcid.org/0000-0002-0529-7048
Yonas E Geda https://orcid.org/0000-0003-1187-7230

Statements and declarations

Author contributions

Jay Shah (Formal analysis; Investigation; Methodology; Visualization; Writing – original draft; Writing – review & editing); Janina Krell-Roesch (Investigation; Methodology; Writing – review & editing); Erica Forzani (Writing – review & editing); David S Knopman (Funding acquisition; Resources; Writing – review & editing); Clifford R Jack (Funding acquisition; Resources; Writing – review & editing); Ronald C Petersen (Funding acquisition; Resources; Writing – review & editing); Yiming Che (Writing – review & editing); Teresa Wu (Conceptualization; Investigation; Methodology; Project administration; Resources; Supervision; Writing – review & editing); Yonas E Geda (Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Writing – review & editing).

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by grants from the National Institutes of Health (NIH) (R01 AG57708). The Mayo Clinic Study of Aging was supported by the National Institutes of Health (U01 AG006786, P30 AG062677, R37 AG011378, R01 AG041851, R01 NS097495), the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic, the Mayo Foundation for Medical Education and Research, the Liston Award, the GHR Foundation, the Schuler Foundation, and used the resources of the Rochester Epidemiology Project (REP) medical records linkage system, which is supported by the National Institute on Aging (NIA: AG 058738), by the Mayo Clinic Research Committee, and by fees paid annually by REP users.

Declaration of conflicting interests

Dr Knopman serves on a Data Safety Monitoring Board for the Dominantly Inherited Alzheimer Network Treatment Unit study. He was an investigator in Alzheimer clinical trials sponsored by Biogen, Lilly Pharmaceuticals and the University of Southern California, both of which have ended, and is currently an investigator in a trial in frontotemporal degeneration with Alector. He has served as a consultant for Roche, AriBio, Linus Health, Biovie and Alzeca Biosciences but receives no personal compensation. He receives funding from the NIH. Dr Jack reports serving on an independent data monitoring board for Roche, has consulted for and served as a speaker for Eisai, and consulted for Biogen, but he receives no personal compensation from any commercial entity. He receives research support from NIH, the GHR foundation, and the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic. Dr Petersen has consulted for Roche, Inc.; Genentech, Inc.; Eli Lilly, Inc.; Nestle, Inc. and Eisai, Inc.; a DSMB for Genentech, Inc. and receives royalties from Oxford University Press for Mild Cognitive Impairment and from UpToDate. His research funding is from NIH/NIA. Dr Geda received funding from Roche, served on the Lundbeck advisory board, and receives research funding from the NIH.
David S Knopman and Teresa Wu are Editorial Board Members of this journal but were not involved in the peer-review process nor had access to any information regarding its peer-review.
The remaining authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

The data used in this study is available to qualified researchers upon reasonable request.

Supplemental material

Supplemental material for this article is available online.

References

1. Lyketsos CG, Lopez O, Jones B, et al. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the Cardiovascular Health Study. *JAMA* 2002; 288: 1475–1483.
2. Geda YE, Roberts RO, Knopman DS, et al. Prevalence of neuropsychiatric symptoms in mild cognitive impairment and normal cognitive aging: population-based study. *Arch Gen Psychiatry* 2008; 65: 1193–1198.
3. Steinberg L. A social neuroscience perspective on adolescent risk-taking. *Dev Rev* 2008; 28: 78–106.
4. Geda YE, Roberts RO, Mielke MM, et al. Baseline neuropsychiatric symptoms and the risk of incident mild cognitive impairment: a population-based study. *Am J Psychiatry* 2014; 171: 572–581.
5. Wilson RS, Krueger KR, Arnold SE, et al. Loneliness and risk of Alzheimer disease. *Arch Gen Psychiatry* 2007; 64: 234–240.
6. Rosenberg PB, Mielke MM, Appleby BS, et al. The association of neuropsychiatric symptoms in MCI with incident dementia and Alzheimer disease. *Am J Geriatr Psychiatry* 2013; 21: 685–695.
7. Palmer K, Di Iulio F, Varsi AE, et al. Neuropsychiatric predictors of progression from amnestic-mild cognitive impairment to Alzheimer's disease: the role of depression and apathy. *J Alzheimers Dis* 2010; 20: 175–183.
8. Pink A, Stokin GB, Bartley MM, et al. Neuropsychiatric symptoms, APOE ϵ4, and the risk of incident dementia: a population-based study. *Neurology* 2015; 84: 935–943.
9. Ramakers IHGB, Visser PJ, Aalten P, et al. Affective symptoms as predictors of Alzheimer's disease in subjects with mild cognitive impairment: a 10-year follow-up study. *Psychol Med* 2010; 40: 1193–1201.
10. Kassem AM, Ganguli M, Yaffe K, et al. Anxiety symptoms and risk of dementia and mild cognitive impairment in the oldest old women. *Aging Ment Health* 2018; 22: 474–482.
11. Krell-Roesch J, Rakusa M, Syrjanen JA, et al. Association between CSF biomarkers of Alzheimer's disease and neuropsychiatric symptoms: Mayo Clinic Study of Aging. *Alzheimers Dement* 2023; 19: 4498–4506.

12. Burhanullah MH, Tschanz JT, Peters ME, et al. Neuropsychiatric symptoms as risk factors for cognitive decline in clinically normal older adults: the Cache County Study. *Am J Geriatr Psychiatry* 2020; 28: 64–71.
13. Ng KP, Chiew H, Rosa-Neto P, et al. Associations of AT(N) biomarkers with neuropsychiatric symptoms in preclinical Alzheimer's disease and cognitively unimpaired individuals. *Transl Neurodegener* 2021; 10: 11.
14. Pink A, Krell-Roesch J, Syrjanen JA, et al. A longitudinal investigation of Aβ, anxiety, depression, and mild cognitive impairment. *Alzheimers Dement* 2022; 18: 1824–1831.
15. Pink A, Krell-Roesch J, Syrjanen JA, et al. Interactions between neuropsychiatric symptoms and Alzheimer's disease neuroimaging biomarkers in predicting longitudinal cognitive decline. *Psychiatr Res Clin Pract* 2023; 5: 4–15.
16. Krell-Roesch J, Syrjanen JA, Vassilaki M, et al. Brain regional glucose metabolism, neuropsychiatric symptoms and the risk of incident mild cognitive impairment: the Mayo Clinic Study of Aging. *Am J Geriatr Psychiatry* 2021; 29: 179–191.
17. Chang C-H, Lin C-H and Lane H-Y. Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease. *Int J Mol Sci* 2021; 22: 2761.
18. Myszczynska MA, Ojamies PN, Lacoste AMB, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol* 2020; 16: 440–456.
19. Falahati F, Westman E and Simmons A. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *J Alzheimers Dis* 2014; 41: 685–708.
20. Meeker KL, Luckett PH, Barthélemy NR, et al. Comparison of cerebrospinal fluid, plasma and neuroimaging biomarker utility in Alzheimer's disease. *Brain Commun* 2024; 6: fcae081.
21. Shah J, Rahman Siddiquee MM, Krell-Roesch J, et al. Neuropsychiatric symptoms and commonly used biomarkers of Alzheimer's disease: a literature review from a machine learning perspective. *J Alzheimers Dis* 2023; 92: 1131–1146.
22. Gill S, Mouches P, Hu S, et al. Using machine learning to predict dementia from neuropsychiatric symptom and neuroimaging data. *J Alzheimers Dis* 2020; 75: 277–288.
23. Roberts RO, Geda YE, Knopman DS, et al. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. *Neuroepidemiology* 2008; 30: 58–69.
24. Kokmen E, Smith GE, Petersen RC, et al. The short test of mental status. Correlations with standardized psychometric testing. *Arch Neurol* 1991; 48: 725–728.
25. Petersen RC. Mild cognitive impairment as a diagnostic entity. *J Intern Med* 2004; 256: 183–194.
26. Rey A. *L'examen clinique en psychologie*. 2e éd. Paris: Presses universitaires de France, 1964.
27. Wechsler D. *WMS-R: Wechsler Memory Scale–Revised: manual*. San Antonio: Psychological Corp.: Harcourt Brace Jovanovich, 1987.
28. Kaplan E, Goodglass H and Weintraub S. *The Boston naming test*. 2nd Edition. Philadelphia: Lea & Febiger, 1983.
29. Lucas JA, Ivnik RJ, Smith GE, et al. Mayo's older Americans normative studies: category fluency norms. *J Clin Exp Neuropsychol* 1998; 20: 194–200.
30. Wechsler D. *WAIS-R: manual : Wechsler adult intelligence scale–revised*. New York, NY: Harcourt Brace Jovanovich [for] Psychological Corp, 1981.
31. Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. *Percept Mot Skills* 1958; 8: 271–276.
32. Jack CR, Wiste HJ, Algeciras-Schimnich A, et al. Predicting amyloid PET and tau PET stages with plasma biomarkers. *Brain* 2023; 146: 2029–2044.
33. Bermudez C, Graff-Radford J, Syrjanen JA, et al. Plasma biomarkers for prediction of Alzheimer's disease neuropathologic change. *Acta Neuropathol* 2023; 146: 13–29.
34. Van Harten AC, Wiste HJ, Weigand SD, et al. CSF Biomarkers in Olmsted County. *Neurology* 2020; 95: e256–e267.
35. Jack CR, Wiste HJ, Weigand SD, et al. Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement* 2017; 13: 205–216.
36. Landau SM, Harvey D, Madison CM, et al. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiol Aging* 2011; 32: 1207–1218.
37. Minoshima S, Frey KA, Foster NL, et al. Preserved pontine glucose metabolism in Alzheimer disease: a reference region for functional brain image (PET) analysis. *J Comput Assist Tomogr* 1995; 19: 541–547.
38. Mielke MM, Dage JL, Frank RD, et al. Performance of plasma phosphorylated tau 181 and 217 in the community. *Nat Med* 2022; 28: 1398–1405.
39. Lowe VJ, Kemp BJ, Jack CR, et al. Comparison of 18F-FDG and PiB PET in cognitive impairment. *J Nucl Med* 2009; 50: 878–886.
40. Jack CR, Lowe VJ, Senjem ML, et al. 11C Pib and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment. *Brain* 2008; 131: 665–680.
41. Kaufer DI, Cummings JL, Ketchel P, et al. Validation of the NPI-Q, a brief clinical form of the neuropsychiatric inventory. *J Neuropsychiatry Clin Neurosci* 2000; 12: 233–239.
42. Beck AT, Steer RA and Brown GK. *BDI-II, Beck depression inventory: manual. Second edition*. San Antonio, TX, Boston: Psychological Corp.; Harcourt Brace, 1996.
43. Beck AT and Steer RA. *BAI: Beck Anxiety inventory: manual*. San Antonio, TX: Psychological Corp., Harcourt Brace, 1993.
44. Müller AC and Guido S. *Introduction to machine learning with Python: a guide for data scientists*. Sebastopol, CA: O'Reilly Media, Inc., 2016.
45. Liashchynskyi P and Liashchynskyi P. Grid search, random search, genetic algorithm: A big comparison for NAS,

http://arxiv.org/abs/1912.06059 (2019, accessed 25 October 2023).
46. Erickson BJ, Korfiatis P, Akkus Z, et al. Machine learning for medical imaging. *Radiographics* 2017; 37: 505–515.
47. Pepe MS, Fan J, Feng Z, et al. The Net Reclassification Index (NRI): a misleading measure of prediction improvement even with independent test data sets. *Stat Biosci* 2014; 7: 282.
48. Hilden J and Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med* 2014; 33: 3405–3414.
49. Demler OV, Pencina MJ and D'Agostino RB. Misuse of DeLong test to compare AUCs for nested models. *Stat Med* 2012; 31: 2577–2587.
50. Mallo SC, Valladares-Rodriguez S, Facal D, et al. Neuropsychiatric symptoms as predictors of conversion from MCI to dementia: a machine learning approach. *Int Psychogeriatr* 2020; 32: 381–392.
51. Almgren H, Camacho M, Hanganu A, et al. Machine learning-based prediction of longitudinal cognitive decline in early Parkinson's disease using multimodal features. *Sci Rep* 2023; 13: 13193.
52. Lundberg S and Lee S-I. A unified approach to interpreting model predictions. arXiv. DOI: 10.48550/arXiv.1705.07874
53. Qiu S, Joshi PS, Miller MI, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain* 2020; 143: 1920–1933.
54. Hüllermeier E, Fober T, Mernberger M, et al. Inductive bias. In: Dubitzky W, Wolkenhauer O and Cho K-H (eds) *Encyclopedia of systems biology*. New York, NY: Springer, 2013, pp.1018–1018.
55. St Sauver JL, Grossardt BR, Leibson CL, et al. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. *Mayo Clin Proc* 2012; 87: 151–160.