Alzheimer's Research & Therapy

RESEARCH

Open Access

Prediction of cognitive conversion within the Alzheimer's disease continuum using deep learning

Siyu Yang[1,2†], Xintong Zhang[3†], Xinyu Du[1], Peng Yan[1], Jing Zhang[1], Wei Wang[2], Jing Wang[4], Lei Zhang[4,5,6,7], Huaiqing Sun[1], Yin Liu[1], Xinran Xu[1], Yaxuan Di[1], Jin Zhong[1], Caiyun Wu[1,8], Jan D. Reinhardt[9,10,11*], Yu Zheng[3*] and Ting Wu[1*]

Abstract

Background Early diagnosis and accurate prognosis of cognitive decline in Alzheimer's disease (AD) is important to timely assignment to optimal treatment modes. We aimed to develop a deep learning model to predict cognitive conversion to guide re-assignment decisions to more intensive therapies where needed.

Methods Longitudinal data including five variable sets, i.e. demographics, medical history, neuropsychological outcomes, laboratory and neuroimaging results, from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort were analyzed. We first developed a deep learning model to predicted cognitive conversion using all five variable sets. We then gradually removed variable sets to obtained parsimonious models for four different years of forecasting after baseline within acceptable frames of reduction in overall model fit (AUC remaining > 0.8).

Results A total of 607 individuals were included at baseline, of whom 538 participants were followed up at 12 months, 482 at 24 months, 268 at 36 months and 280 at 48 months. Predictive performance was excellent with AUCs ranging from 0.87 to 0.92 when all variable sets were considered. Parsimonious prediction models that still had a good performance with AUC 0.80–0.84 were established, each only including two variable sets. Neuropsychological outcomes were included in all parsimonious models. In addition, biomarker was included at year 1 and year 2, imaging data at year 3 and demographics at year 4. Under our pre-set threshold, the rate of upgrade to more intensive therapies according to predicted cognitive conversion was always higher than according to actual cognitive conversion so as to decrease the false positive rate, indicating the proportion of patients who would have missed upgraded treatment based on prognostic models although they actually needed it.

†Siyu Yang and Xintong Zhang contributed equally to this work.

*Correspondence:
Jan D. Reinhardt
reinhardt@scu.edu.cn
Yu Zheng
zhengyu8710@163.com
Ting Wu
wuting80000@126.com

Full list of author information is available at the end of the article

Conclusions Neurophysiological tests combined with other indicator sets that vary along the AD continuum can improve can provide aid for clinical treatment decisions leading to improved management of the disease.

Trail registration information ClinicalTrials.gov Identifier: NCT00106899 (Registration Date: 31 March 2005).

Keywords Alzheimer's disease, Machine learning, Cognitive conversion, Prediction model

Background

In 2018, the US National Institute on Aging and Alzheimer's Association, using amyloidosis, tau pathology and neurodegeneration (ATN), redefined Alzheimer's disease (AD) moving from a syndromal to a biological construct [1], thus allowing clinicians and researchers to better delineate different phases of clinical disease progression including preclinical and prodromal AD [2, 3]. The current treatment algorithm foresees initial treatment with low dose cholinesterase inhibitors (ChEIs) followed by an increase of the ChEIs dose and switching to memantine as AD progresses. With further progression, combination therapy is recommended or other treatment modalities are sought, for example, immunotherapy targeting β amyloid (Aβ) and tau [4–8].

As initial treatments can often not effectively halt or slow down disease progression in individual patients, it is important to predict patient response based on available patient data and clinical information in order to make early upgrade decisions. A problem herein is that the operationalization of disease progression is complex involving a variety of cognitive tests, plasma and cerebrospinal fluid (CSF) biomarkers, and radiological imaging [9–11].

A number of studies has, therefore, applied machine learning algorithms using high dimensional data combining comprehensive information from the above sources to predict disease progression in AD. Since advanced neuroimaging such as amyloid positron emission tomography (PET) or tau PET is not readily available in routine clinical practice due to cost and radioactive burden on the patient and the extraction of CSF sampling is invasive [12, 13], prediction algorithms using easily available plasma and cerebrospinal biomarkers such as plasma A β 42/Aβ 40, phosphorylated tau (p-tau) and neurofilament light (NfL), routine imaging data, and information from cognitive tests such as mini mental state examination (MMSE), Alzheimer's disease assessment scale-cognition (ADAS-cog) and auditory verbal learning test (AVLT) may be more suitable to clinical practice, particularly in lower resourced settings [14–20]. Moreover, defining parsimonious sets for prediction models will increase applicability in the clinical context.

In this study, we used information from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database including demographic data, genetic genotype, biomarkers, neuropsychology tests and neuroimaging to select variable sets and develop prediction models for AD disease progression on which upgrade of treatment decisions can be based. Our specific aims were (1) to develop machine learning models to predict cognitive conversion with few and accessible indicators; (2) to compare the accuracy of models using different sets of such predictors; (3) make recommendations for implementation of such algorithms in clinical practice.

Methods

Study design

This is a modelling study based on prospective cohort data extracted from the ADNI database. The original study from the ADNI is a multicenter study aimed at early detecting and stopping the progression of AD with data being collected since 2004 [21].

Setting

Individual patient data from the ADNI database were included in our study if the following information was available: (1) plasma and CSF biomarker; (2) baseline and longitudinal neuropsychological assessments; (3) average thickness of the middle temporal lobe; (4) Apolipoprotein E (APOE) genotyping. For each patient, the first available data point served as baseline in our study. We chose 12-month spacing between time points based on frequency of follow-up visits to ensure efficient data for model building (baseline, 12 months, 24 months, 36 months and 48 month).

Participants

The original ADNI study included participants with cognitive normal state (CN), subjective cognitive decline (SCD), mild cognitive impairment (MCI) and AD. Detailed eligibility criteria are available from URL: www.adni-info.org. Criteria for the classification of subjects into different phases of AD progression are provided in Table S1.

Data sources/measurement

Demographics

The following demographic data were assessed by questionnaire: age, gender, education, race, marriage, treatment and medical history.

APOE genotyping

APOE genotyping was performed in all participants with polymerase chain reaction (PCR) following the Hixson and Vernier protocol, and the test was considered positive if one or more $\varepsilon 4$ allele was detected ($\varepsilon 4+$).

Neuropsychological assessment

Subjects were evaluated with the following tests: Hachinski ischemic scale (HIS) [22], MMSE [23], ADAS-cog) Montreal cognitive assessment (MOCA) [24], auditory verbal learning test including immediate recall (AVLT-IM), learning (AVLT-L), forgetting (AVLT-IF) and percent forgetting (AVLT-PC) [25], clinical dementia rating (CDR) [26], neuropsychiatric inventory (NPI) [27], geriatric depression scale (GDS) [28], functional assessment questionary (FAQ) [29], logical memory-delayed recall (LDEL) [30] and Trail Making Test B (TRA B) [31].

Plasma and CSF biomarker measurements

Plasma p-tau181 and NfL were analyzed with single molecule array (Simoa) [32], using in-house assays developed in the Clinical Neurochemistry laboratory, University of Gothenburg, Sweden, with the assay for p-tau181 being based on a combination of two monoclonal antibodies (Tau12 and AT270) to measure N-terminal versus mid-domain forms of p-tau181, and NfL measurement using a combination of monoclonal antibodies and purified bovine NfL as a calibrator.

Measurements of concentrations of CSF $A\beta_{1-42}$, t-tau, p-tau181 were made with micro-bead-based multiplex immunoassay, and the INNO-BIA AlzBio3 RUO test (Fujirebio, Ghent, Belgium) on the Luminex platform.

Structural MRI analyses

Subjects underwent a 3-Tesla magnetic resonance imaging (MRI) scan of the brain. Cortical thickness of the middle temporal brain region was measured using FreeSurfer (version 4.3).

Study size

We did not employ statistical methods to determine the sample size. Instead, we included 607 participants by utilizing all the available data from the ADNI database that had baseline plasma biomarker information and follow-up data from all four timepoints.

Variables

Predictors used in our machine learning models included information on demographics, medical history, neuropsychological outcomes, and laboratory and neuroimaging results with details provided in supplementary Table S2. Time-dependent variables such as ADAS-cog, MMSE, CDR, MOCA, GDS, NPI, LDEL, plasma p-tau181, NfL and average thickness of the middle temporal lobe for the evaluation of neurobehavioral status were recorded in 12-months intervals.

Following the guidelines published by the American College of Physicians and the American Academy of Family Physicians [33], we used the ADAS-cog score change as the primary outcome of our study. Individuals who had an improvement of 4 or more points on the ADAS-cog were considered to have cognition improved (CoI), while the others were classified as cognition not improved (CNI).

Statistical methods

For data pre-processing, one-hot encoding was first employed to transform categorical features into a binary format where only one bite is 1 and the rest are 0. We opted for one-hot encoding because it effectively handles unordered categorical data. This approach prevents bias that could arise from introducing irrelevant numerical relationships. To address missing values, we imputed a constant value of 255 for time-series variables, which was not factored into the supervised learning process, while for other variables containing missing data, we utilized the k-nearest neighbors (KNN) algorithm for imputation. This technique estimates missing values by identifying the. K most similar samples using Euclidean distance and imputing missing values based on the mean (for continuous variables) or mode (for categorical variables) of these neighbors. In our study, we set K=10 with equal weights for all neighbors. This non-parametric approach leverages data similarity to provide robust imputations, enhancing data completeness and supporting the reliability of subsequent machine learning analyses. Outliers were identified and removed with box-whisker plots. Specifically, outliers were identified using the Interquartile Range (IQR) rule, defined as follows: Lower bound: Q1-1.5×IQR, Upper bound: Q3-1.5×IQR. Q1 and Q3 represent the 25th and 75th percentiles, respectively, and IQR = Q3 – Q1. Data points outside this range were classified as outliers and processed accordingly. For standardized scaling, min-max normalization was carried out for both ordinal and quantitative variables.

Potential predictors were categorized into five sets (demographic characteristics, genetic features, neuropsychological test, plasma biomarkers and MRI measure), yielding a total of 31 unique combinations (one combination with five sets, five combinations with four sets, ten with three sets, 10 with two groups, and five with one set). Prediction models of cognitive conversion at four time points (1-year, 2-year, 3-year, and 4-year) were developed for each combination of sets using automated machine learning (AutoGluon version 0.3.1 [34]), including ten algorithms (LightGBM, CatBoost, XGBoost, Random Forest, Extremely Randomized Trees, K-nearest neighbors, Linear Regression, Neural Network

Implemented in MXNet, Network and Neural Network with FASTAI backend). We performed five-fold cross-validation to test the model stability and to obtain the predicted outcomes of all patients. Five-fold cross-validation, a well-established and rigorous approach widely used in machine learning for model evaluation. This method involves dividing the dataset into five subsets of equal size, where each subset serves as a validation set once, while the remaining four subsets are used for training. This process is repeated five times, ensuring that every data point is used both for training and validation. The primary advantage of five-fold cross-validation lies in its ability to mitigate bias and variance. By averaging the results across all five iterations, it provides a robust and stable estimate of the model's generalization performance, reducing the risk of overfitting or relying on any specific data split. This approach maximizes the utilization of the available dataset and offers a comprehensive assessment of model performance across different subsets, thereby enhancing the reliability of the evaluation. Area under curve value (AUC) was used to evaluate model performance, with 95% confidence intervals being calculated employing bootstrap with 2000 replicates. AUC values range from 0.5 to 1.0, with 0.5 denoting random guess and 1.0 perfect prediction. AUC value of 0.80 was used as cutoff to identify set combinations that maintained good predictive performance while reducing the number of predictor sets at each time point. In order to minimize false positives, a cognitive improvement of 4 on the ADAS-cog scale was set as cutoff for upgrade to more aggressive therapy mode. Confusion matrices were further provided which comprises of four components namely true positive (TP), true negative (TN), false negative (FN) and false positive (FP). We also calculated the area under the precision-recall curve (AUPRC), a commonly used metric in imbalanced data to measure the model's ability to identify rare events. Other performance measures included accuracy, sensitivity/recall, specificity, positive predictive value (PPV), negative predictive value (NPV) with 95% confidence interval (95% CI) and F_β scores. The following expressions are employed for the computation of metrics:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

F_β scores were calculated according to the following equation: $F_\beta=(1+\beta^2) \times \frac{\text{presion} \times \text{recall}}{\beta^2 \times \text{presion}+\text{recall}}$, β was set to 0.5 as to improve recall. Optimal cut-off values were determined through largest $F_{0.5}$ scores, indicating low false positives and negatives. All statistical analyses were conducted with Python 3.9 and STATA 16.0 (StataCorp LLC, TX, United States). Statistical testing was two-tailed, and level of statistical significance was set at alpha=0.05.

Results

Demographics characteristics by cognitive status

Patients' overall demographic characteristics and distributed according to cognitive improvement (i.e., improved or not improved) at 12, 24, 36 and 48 months are provided in Supplementary Table S3. From the 607 individuals included at baseline, 538 participants were followed at 12 months, 482 at 24 months, 268 at 36 months and 280 at 48 months. The flowchart of the screening process was showed in Figure S1. One hundred and sixty patients at 12 months, 152 at 24 months, 105 at 36 months and 92 at 48 months were classified as CoI, whereas 567 at 12 months, 506 at 24 months, 272 at 36 months and 311 at 48 months were considered as CNI.

Predictive performance of automatic machine learning modeling

Figure 1 summarizes the model selection process and main results.

Predictor sets considered in the models included demographic data (demographic set-d), genetic data (genetic set-g), tests for cognitive function (cognitive set-c), plasma biomarkers (biomarker set-b) and MRI measure (imaging set-i).

We first selected models with highest AUC indicating best model fit. This were models featuring all five variables sets at all timepoints (AUC = 0.87 at 12 months, AUC = 0.87 at 24 months, AUC = 0.90 at 36 months, AUC = 0.92 at 48 months). Successively dropping variable sets while keeping the model with lowest number of variable sets that achieved AUC above 0.8 at the same time, we identified parsimonious models. For each timepoint, these were models including two predictors sets: group cb at 12 months, AUC = 0.82; group cb at 24 months, AUC = 0.80; group ci at 36 months, AUC = 0.84; group dc at 48 months, AUC = 0.80). If only one variable set was considered, models including the cognitive set al.ways achieved the best fit, however always below AUC 0.8. Detailed information on the models selected in each information reduction step is provided in Figs. 2 and 3, Supplementary Figure S2-S3 and Table S4-S5. Figure 4 showed AUCs given with 84%CIs so that non-overlap of

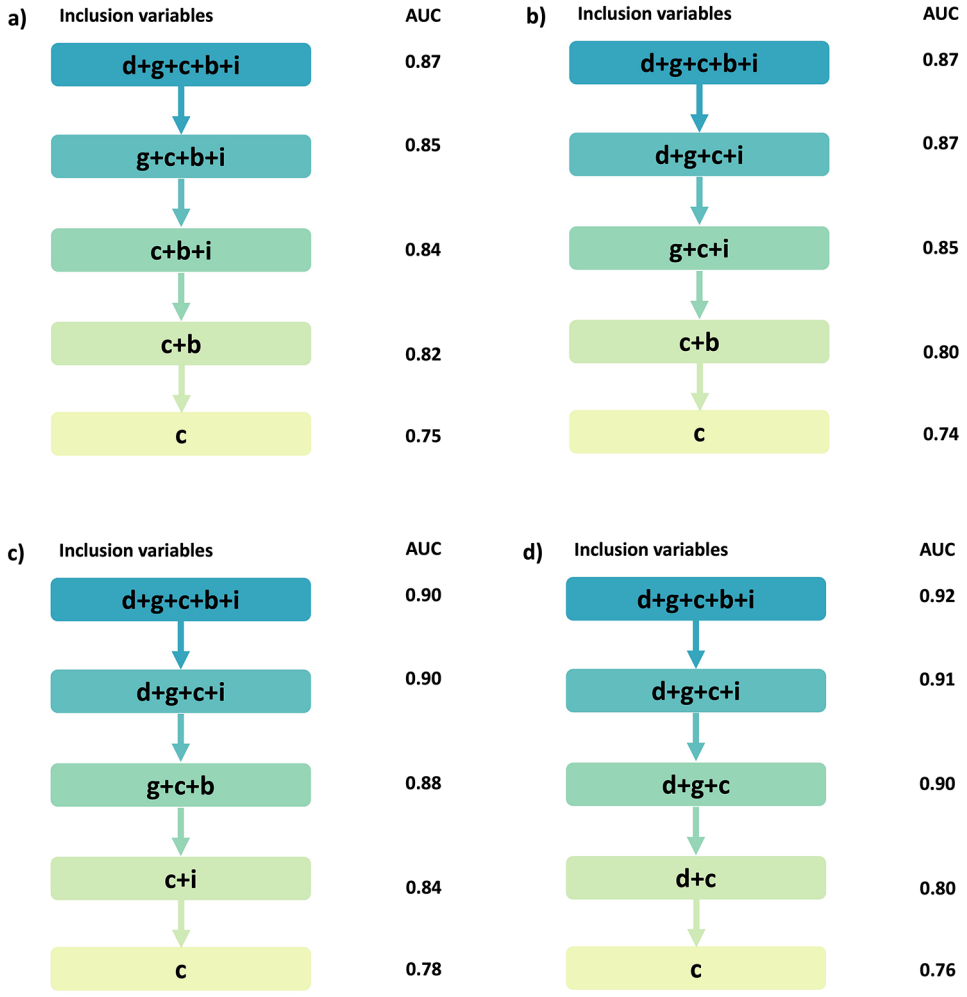

Fig. 1 Feature selection process in prediction of cognitive improvement. (**a**) 12 months; (**b**) 24 months; (**c**) 36 months; (**d**) 48 months. Notes: demographic data-d; genetic data-g; cognitive data-c; biomarker-b; imaging data-i

CIs for different models indicates that the difference is approximately statistically significant at $p < 0.05$ [35].

In addition, we further analyzed the performance of the model by subgroup according to diagnosis. In best fitting models, the AUCs of the CN group exceeded 0.9 across all time points. Notably, the SCD group achieved a peak AUC of 0.940 (95% CI: 0.790-1.000) at the 12-month, whereas MCI group recorded the lowest AUC of 0.769 (95% CI: 0.680–0.857). For the AD group, the AUC peaked at 0.961 (95% CI: 0.859-1.000) at 12-month but subsequently declined. However, data for the AD group at 36 and 48 months were unavailable. Detailed statistical details can be found in Supplementary Table S6. Considering parsimonious models, the AUC values for the CN group consistently exceeded 0.8 across all time points. In contrast, the AUCs for the AD group remained around 0.5 at 12-month and 24-month. SCD group demonstrated the highest AUC at 12 months, while the MCI group showed a modest increase over time, surpassing 0.7 after 12-month. Detailed statistical details are presented in Table S7.

Proposed treatment upgrade following actual and predicted cognitive improvement at 12, 24, 36 and 48 months

Proportions of patients who needed upgraded treatment as determined by suboptimal cognitive improvement were 85.32% (459/538) at 12 months, 79.67% (384/482) at 24 months, 75.37% (202/268) at 36 months, and 76.43% (214/280) at 48 months in the actual data.

Considering the overall best fitting models (i.e. those which included all variable sets), predicted conversion resulted in upgraded treatment for 90.15% (485/538) of patients at 12 month, 90.25% (435/482) at 24 month, 81.72% (219/268) at 36 months, and 82.86% (232/280) at 48 months. Detailed statistics regarding subgroup analysis of treatment upgrade are provided in Table 1; Fig. 5.

As regards parsimonious models, predicted cognitive change that would indicate early treatment upgrade

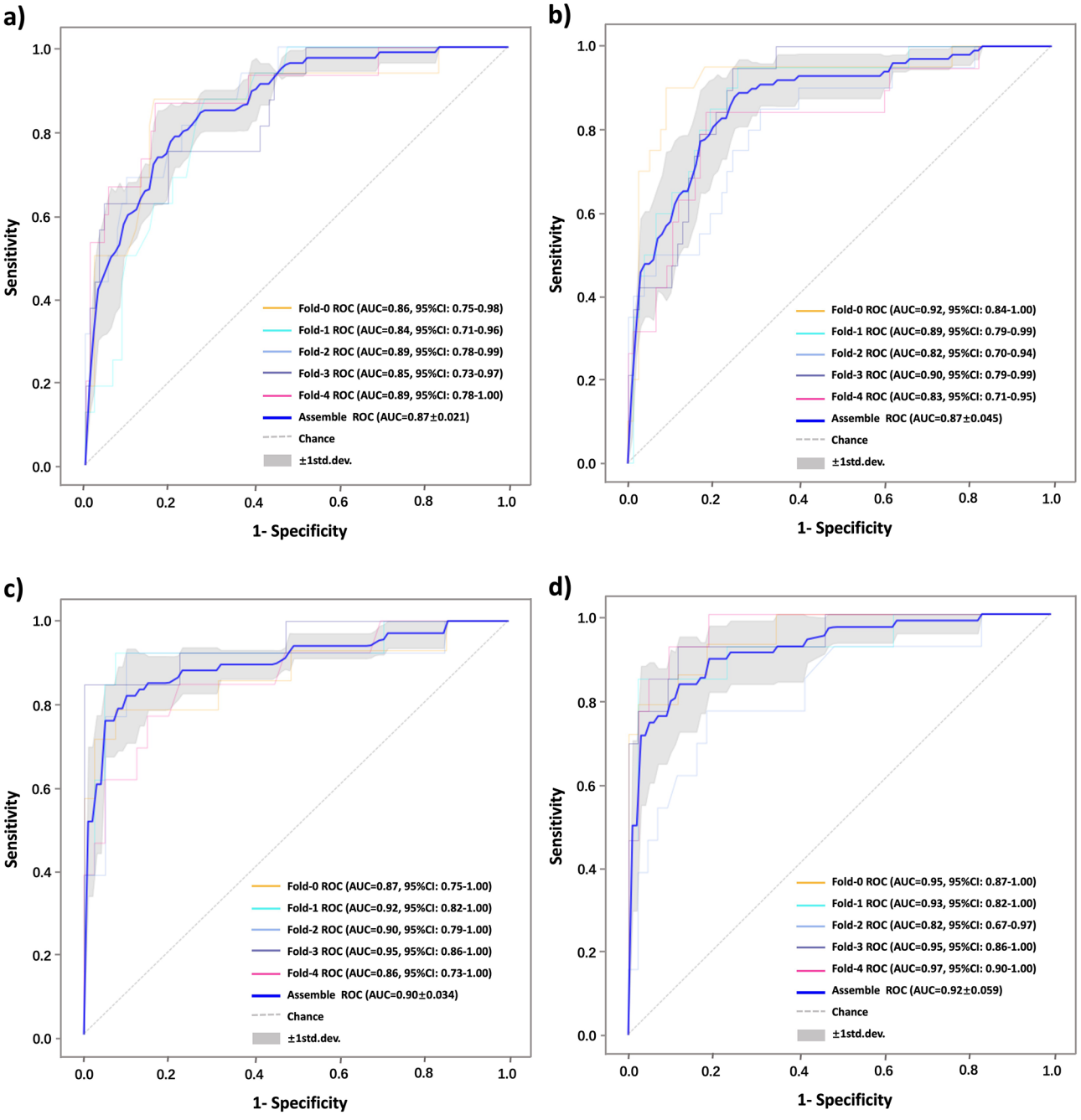

Fig. 2 ROC comparisons of cognitive improvement at 12, 24, 36 and 48 months with best model fit. (**a**) 12 months; (**b**) 24 months; (**c**) 36 months; (**d**) 48 months

in 90.15% (485/538) of the patients at 12 month, 89.21% (430/482) at 24 month, 82.46% (221/268) at 36 months, and 79.29% (222/280) at 48 months. Detailed statistics regarding subgroup analysis of treatment upgrade are provided in Table 1; Fig. 6.

Further subgroup analysis based on different diagnoses showed that the proportion of upgraded treatment was quite high in the early stage of AD and MCI groups. Detailed statistics are provided in Table S8 and Table S9.

False positive rate of different models at 12, 24, 36 and 48 months

Table 2 summarizes false positive rates (FPR) that is the proportion of patients who would have missed upgraded treatment based on the prognostic models although they actually needed it. The average false positive rate across five folds was acceptable in different models at four time-points except for set combination dc where the FPR

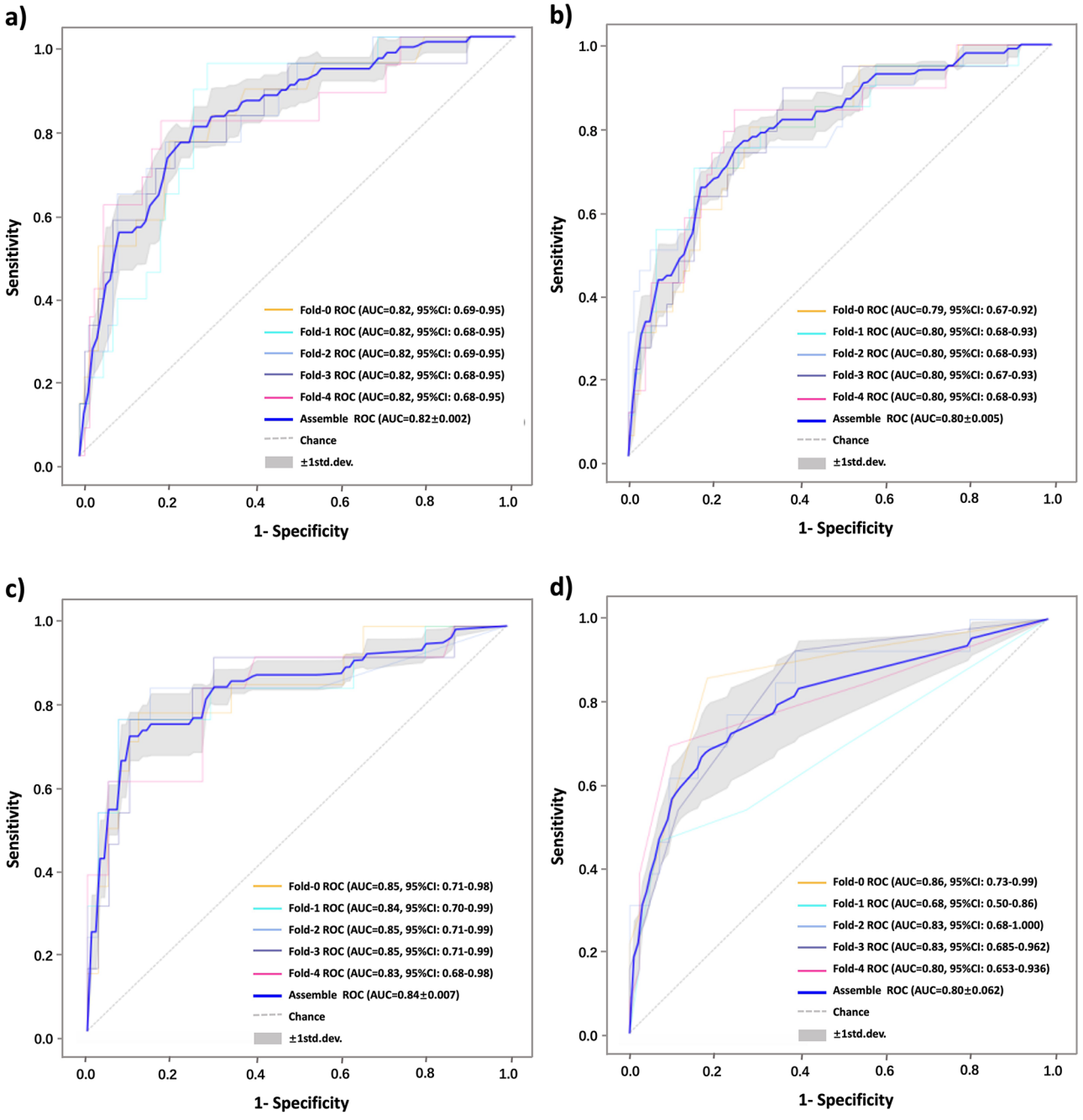

Fig. 3 ROC comparisons of cognitive improvement at 12, 24, 36 and 48 months with parsimonious models. (**a**) 12 months; (**b**) 24 months; (**c**) 36 months; (**d**) 48 months

reached 9.35%. Other models stabilized over time at 48 months.

Subgroup analysis according to diagnosis

In best fitting models, the AUCs of the CN group exceeded 0.9 across all time points. Notably, the SCD group achieved a peak AUC of 0.940 (95% CI: 0.790-1.000) at the 12-month, whereas MCI group recorded the lowest AUC of 0.769 (95% CI: 0.680–0.857). For the AD group, the AUC peaked at 0.961 (95% CI: 0.859-1.000) at 12-month but subsequently declined. However, data for the AD group at 36 and 48 months were unavailable. Detailed statistical details can be found in Supplementary Table S6.

Considering parsimonious models, the AUC values for the CN group consistently exceeded 0.8 across all time points. In contrast, the AUCs for the AD group remained around 0.5 at 12-month and 24-month. SCD group

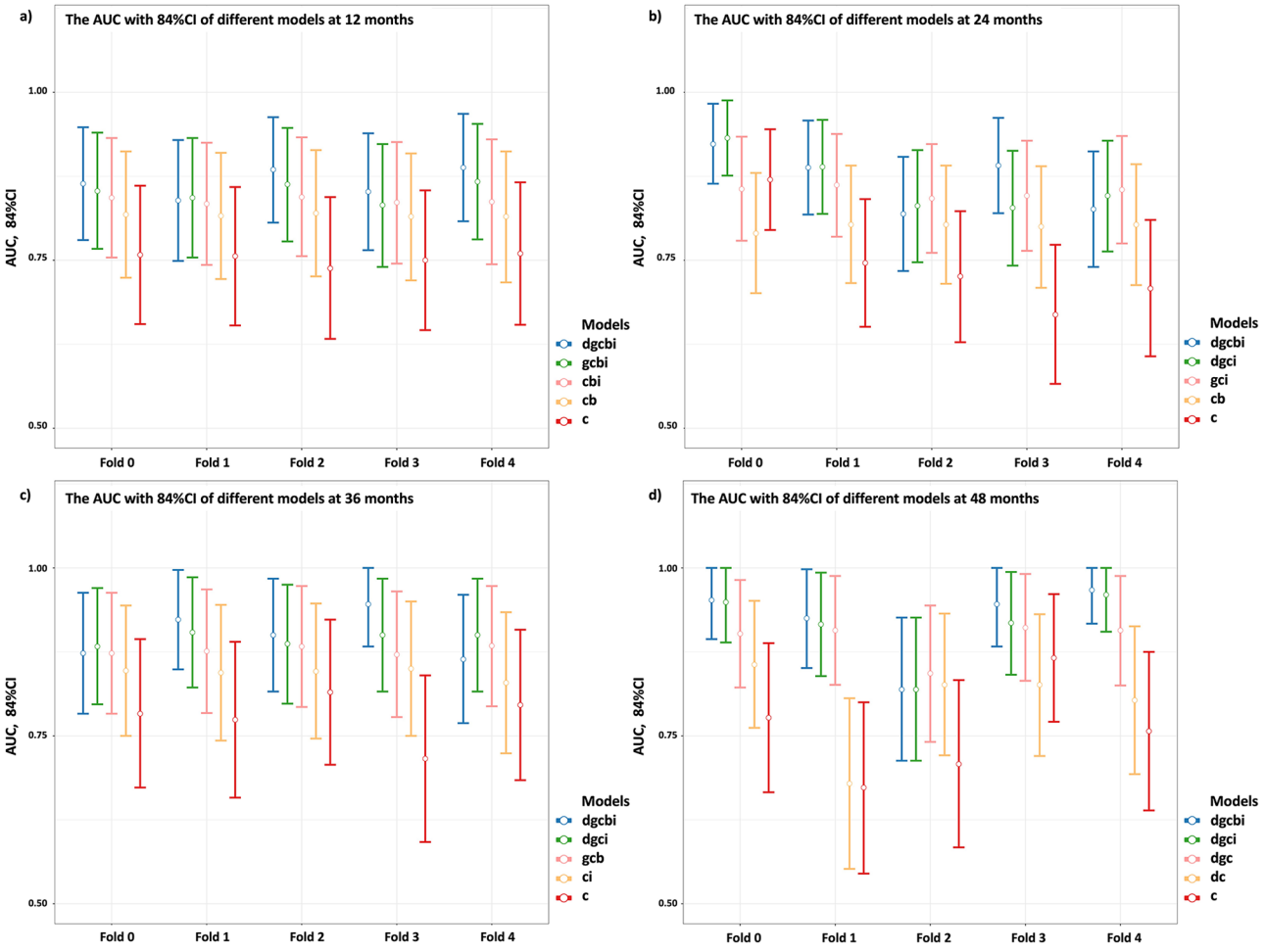

Fig. 4 The AUCs with 84% CI of different models. (**a**) 12 months; (**b**) 24 months; (**c**) 36 months; (**d**) 48 months. Notes: demographic data-d; genetic data-g; cognitive data-c; biomarker-b; imaging data-i

demonstrated the highest AUC at 12 months, while the MCI group showed a modest increase over time, surpassing 0.7 after 12-month. Detailed statistical details are presented in Table S7.

For asymptomatic patients, the early use of the scale is simple to operate and can effectively monitor the dynamic changes in cognitive levels.

Discussion

In this study, we developed and tested a model using deep learning algorithms to predict changes in cognitive function of AD patients based on various combinations of different variable sets. Predictive performance was excellent with AUCs ranging from 0.87 to 0.92 when all variable sets were considered. Parsimonious prediction models that still had a good predictive performance of 0.80–0.84 could also be established, each only including two variable sets. This is important as in practice not all relevant 31 variables included in the sets can be easily collected or at least absorb considerable assessment time. In particular, parsimonious models also achieved low false positive rates, that is proportions of patients who should have been assigned to upgraded treatment because of suboptimal cognitive development but would not have been based on models' predictions were kept low.

Longitudinal data results revealed that the predictive performance of our algorithm improved over time. Moreover, we progressively filtered the full set of combined variables to reduce variable set combinations to achieve parsimony while keeping acceptable predictive performance. The results showed that at least two variable combinations were needed to achieve satisfactory AUCs. Importantly, we found that in parsimonious models the neuropsychological variable set was always included. In single-variable combinations, the cognitive scale also achieved the highest predictive performance, with AUC being above 0.7 at four time points. In practice, this is easy to implement because neuropsychological scales can be assessed without great amounts of resources and time.

Table 1 Comparison between actual and AI-predicted results of full and parsimonious models after 12, 24, 36 and 48 months

		Actual reassignment status	Predicted reassignment status	
			Full model	Parsimonious model
12-month	**Overall, n (%)**	459/538 (85.32)	485/538 (90.15)	485/538 (90.15)
	Observation	312/380 (82.11)	334/380 (87.89)	334/380 (87.89)
	Monotherapy-ChEI	93/101 (92.08)	97/101 (96.04)	95/101 (94.06)
	Monotherapy-Memantine	49/50 (98.00)	48/50 (96.00)	49/50 (98.00)
	Combined therapy	5/7 (71.43)	6/7 (85.71)	7/7 (100.00)
24-month	**Overall, n (%)**	384/482 (79.67)	435/482 (90.25)	430/482 (89.21)
	Observation	304/386 (78.76)	341/386 (88.34)	339/386 (87.82)
	Monotherapy-ChEI	54/62 (87.10)	61/62 (98.39)	59/62 (95.16)
	Monotherapy-Memantine	25/31 (80.65)	31/31 (100.00)	31/31 (100.00)
	Combined therapy	1/3 (33.33)	2/3 (66.67)	1/3 (33.33)
36-month	**Overall, n (%)**	202/268 (75.37)	219/268 (81.72)	221/268 (82.46)
	Observation	157/198 (79.29)	170/198 (85.86)	169/198 (85.35)
	Monotherapy-ChEI	36/50 (72.00)	39/50 (78.00)	38/50 (76.00)
	Monotherapy-Memantine	7/17 (41.18)	10/17 (58.82)	12/17 (70.59)
	Combined therapy	2/3 (66.67)	0/3 (0.00)	2/3 (66.67)
48-month	**Overall, n (%)**	214/280 (76.43)	232/280 (82.86)	222/280 (79.29)
	Observation	183/233 (78.54)	199/233 (85.41)	187/233 (80.26)
	Monotherapy-ChEI	26/35 (74.29)	26/35 (74.29)	27/35 (77.14)
	Monotherapy-Memantine	5/10 (50.00)	7/10 (70.00)	7/10 (70.00)
	Combined therapy	0/2 (0.00)	0/2 (0.00)	1/2 (50.00)

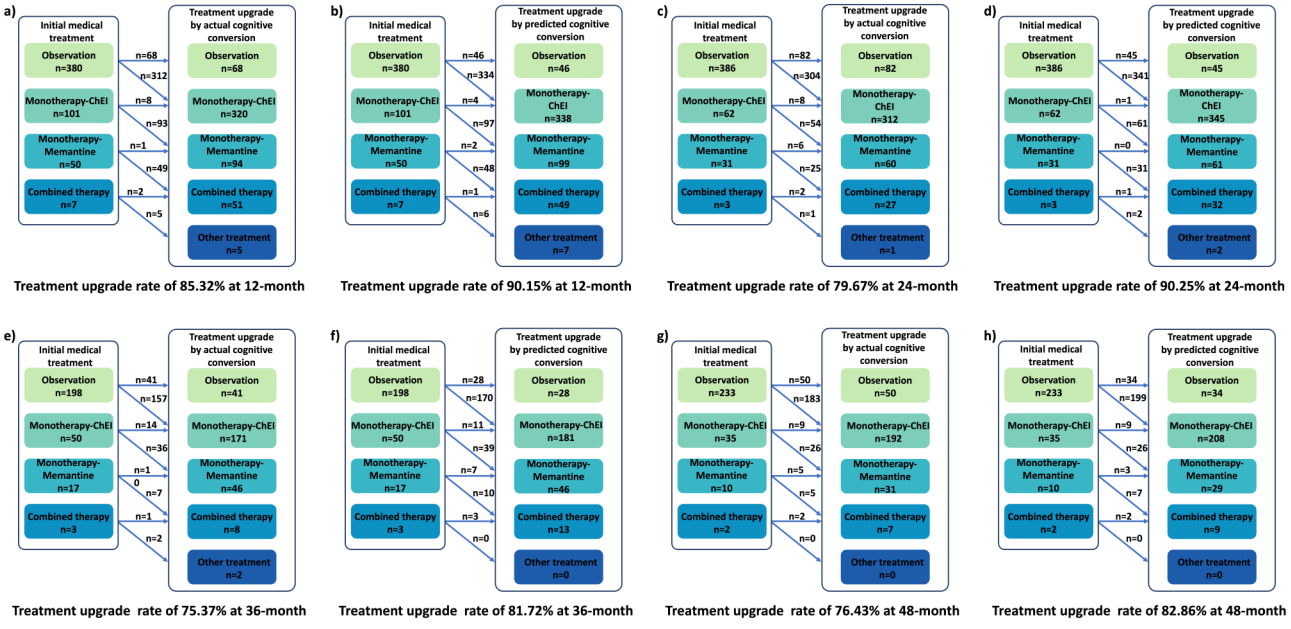

Fig. 5 Proposed treatment upgrade of best model fit following actual and predictive cognitive improvement at 12, 24, 36 and 48 months. (**a**) Treatment upgrade following actual cognitive improvement at 12-month; (**b**) Treatment upgrade following predictive cognitive improvement 12-month; (**c**) Treatment upgrade following actual cognitive improvement at 24-months; (**d**) Treatment upgrade following predictive cognitive improvement at 24-month; (**e**) Treatment upgrade following actual cognitive improvement at 36-month; (**f**) Treatment upgrade following predictive cognitive improvement at 36-month; (**g**) Treatment upgrade following actual cognitive improvement at 48-month; (**h**) Treatment upgrade following predictive cognitive improvement at 48-month

We also saw that second most important sets for prediction in parsimonious models varied at different time points, with biomarkers playing an important role at the first two years of forecast, then imaging results at the third year, and finally demographic data. This is in line with clinical data showing the presence of biomarkers such as plasma Aβ and tau in AD patients a decade or two prior to the manifestation of clinical symptoms.

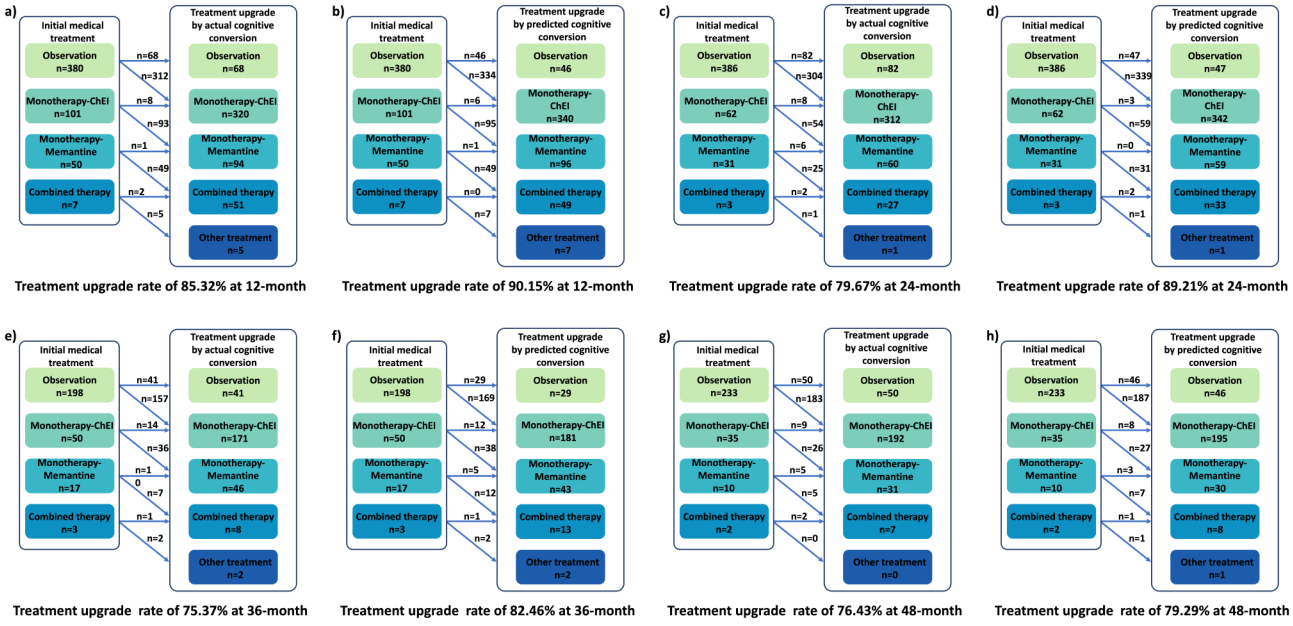

Fig. 6 Proposed treatment upgrade of parsimonious models following actual and predictive cognitive improvement at 12, 24, 36 and 48 months. (**a**) Treatment upgrade following actual cognitive improvement at 12-month; (**b**) Treatment upgrade following predictive cognitive improvement 12-month; (**c**) Treatment upgrade following actual cognitive improvement at 24-months; (**d**) Treatment upgrade following predictive cognitive improvement at 24-month; (**e**) Treatment upgrade following actual cognitive improvement at 36-month; (**f**) Treatment upgrade following predictive cognitive improvement at 36-month; (**g**) Treatment upgrade following actual cognitive improvement at 48-month; (**h**) Treatment upgrade following predictive cognitive improvement at 48-month

With disease progression, patients commence experiencing alterations in neuroimaging, predominantly characterized by temporal lobe atrophy. Consequently, the incorporation of MRI measurements during this stage exhibits notable predictive capability. In the advanced stages, owing to the clinical symptom heterogeneity, limited efficacy of pharmaceutical interventions and stabilization of biomarker level, alternative predictors demonstrate inferior performance compared to demographic data, which effectively reflect the patient's social status, cognitive reserve, and possibly caregiver availability and support.

We further analyzed predictive outcomes of models according to a pre-set threshold for upgrade to more intensive therapy regimen, keeping the conversion rate predicted by the deep learning model slightly higher than actual results at various time points. This ensures that no patient in need of an upgraded treatment is overlooked. Only at 48 months did the parsimonious model exhibit a higher FPR, suggesting that there were predictions of improvement for patients who, in reality, did not experience such improvements, subsequently leading to delayed upgrades in their treatment regimen. This is possible owing to missing predictors such as environmental factors. This was less pronounced when the model including all sets was used. On the other hand, the full model exhibited a decrease in FPR from 24 to 48 months, while the parsimonious model showed a stable FPR at all timepoint except for 48 months. These findings warrant further investigation to better comprehend the underlying factors influencing model performance over time. The identification of factors leading to false positives can have significant implications for patient treatment strategies and overall clinical decision-making. Further research is needed to elucidate the mechanisms contributing to the observed trends in FPR and to enhance the predictive accuracy of such models.

The only previous study that considered all variable sets included in our study did not predict cognitive improvement but progression to AD. The authors found that combing plasma biomarker, memory, executive function and APOE produced the highest prediction accuracy (AUC = 0.91) which still remained good with plasma p-tau217 (AUC = 0.83) [36].

Ocasio and colleagues [37] developed a CNN network to predict MCI conversion to AD at three years using longitudinal and whole-brain 3D MRI. There results showed an accuracy of 0.793 with the most important regions including lateral ventricles, periventricular white matter and cortical gray matter. The parsimonious models identified in our study have similar or better predictive capacity.

Our study has important clinical implications. Accurate diagnosis and timely intervention at the preclinical and prodromal stages of AD have become core aims of drug development [38]. To optimize therapy assignment,

Table 2 False positive rate of different models at 12, 24, 36 and 48 months

	False positive rate				
12-month	**dgcbi**	**gcbi**	**cbi**	**cb**	**c**
Fold0	2.17%	3.26%	2.17%	4.35%	1.09%
Fold1	9.78%	0.00%	10.87%	0.00%	6.52%
Fold2	0.00%	3.26%	9.78%	8.70%	13.04%
Fold3	3.26%	4.35%	0.00%	2.17%	5.43%
Fold4	1.10%	1.10%	4.40%	5.49%	1.10%
Average across folds	3.26%	2.39%	5.44%	4.14%	5.44%
24-month	**dgcbi**	**dgci**	**gci**	**cb**	**c**
Fold0	2.60%	3.90%	11.69%	2.60%	10.39%
Fold1	2.60%	7.79%	5.19%	6.49%	5.19%
Fold2	0.00%	1.30%	2.60%	1.30%	0.00%
Fold3	1.30%	0.00%	1.30%	2.60%	0.00%
Fold4	0.00%	1.32%	10.53%	5.26%	6.58%
Average across folds	1.30%	2.86%	6.26%	3.65%	4.43%
36-month	**dgcbi**	**dgci**	**gcb**	**ci**	**c**
Fold0	0.00%	4.88%	0.00%	7.32%	4.88%
Fold1	4.88%	0.00%	2.44%	2.44%	2.44%
Fold2	5.oo%	2.50%	2.50%	2.50%	7.50%
Fold3	0.00%	0.00%	0.00%	10.00%	0.00%
Fold4	0.00%	2.50%	5.00%	0.00%	17.50%
Average across folds	1.98%	1.98%	1.99%	4.45%	6.46%
48-month	**dgcbi**	**dgci**	**dgc**	**dc**	**c**
Fold0	0.00%	0.00%	0.00%	18.60%	4.65%
Fold1	2.33%	2.33%	2.33%	6.98%	6.98%
Fold2	2.33%	0.00%	0.00%	0.00%	4.65%
Fold3	0.00%	0.00%	0.00%	11.63%	0.00%
Fold4	2.38%	2.50%	2.38%	9.52%	2.38%
Average across folds	1.41%	0.97%	0.94%	9.35%	3.73%

Notes: demographic data-d; genetic data-g; cognitive data-c; biomarker-b; imaging data-i

we need to timely identify individuals who are at a high risk of developing AD. While presently, many studies use AD prediction models that differentiate between NC and AD, MCI and AD, our model predicts annual cognitive changes for both asymptomatic and symptomatic individuals. Prediction of cognitive progression a year later with respective adjustment of treatments appears more important in clinical settings then the prediction of change in diagnosis. The performance of our prediction model at various time points has significant implications for optimizing treatment escalation strategies. We established a prediction threshold to identify patients requiring more intensive interventions early. To minimize the risk of missed diagnoses, we set the threshold conversion rate predicted by the deep learning model slightly higher than observed clinical data, ensuring timely intervention for patients needing enhanced care. Future studies should aim to optimize the decision threshold to balance sensitivity and specificity, possibly by dynamically adjusting it based on patient history and disease progression. Implementing predictive models in clinical practice requires evaluating the cost-effectiveness of included variables and their impact on healthcare resources. Our findings indicate that neuropsychological assessments play a pivotal role in the parsimonious model, consistently achieving an AUC above 0.7 across all time points. Neuropsychological assessments are cost-effective and time-efficient, making them valuable tools for early detection without imposing significant medical burdens. Other important variables, such as biomarkers, neuroimaging, and demographics, vary in relevance depending on the disease stage. A phased clinical implementation strategy is proposed: in the early stage, prioritize neuropsychological assessments and selectively apply biomarker tests for high-risk patients to balance cost and effectiveness. In the medium stage, neuroimaging can be selectively used for patients exhibiting cognitive decline, optimizing both diagnostic accuracy and cost-efficiency. In the later stages, demographic and neuropsychological data should guide interventions, reducing reliance on expensive biomarkers and imaging. This staged approach maximizes the predictive utility of different variables at each disease phase, enhancing cost-effectiveness while maintaining high accuracy.

Limitations

This study has some limitations that warrant mentioning. First, our sample size was relatively small and generalizability is limited as data are mostly from the USA. In addition loss to follow up is problem as it may be related to cognitive conversion and measured and unmeasured influence factors. Unfortunately, longitudinal multiple imputation models for deep learning are still unsatisfactory [39]. Second, due to the relatively small sample size we did not perform subgroup analysis. Therefore, as of yet, no data indicating difference in model performance across different subgroups are available. Third, we used F-score instead of Youden index as the cut-off value and other thresholds according to the demand of clinical settings may be desirable. Eventually, our study lacks an external validation cohort., and caution is thus warranted when interpreting our results. We acknowledge the importance of external validation to further confirm the model's applicability in real-world scenarios. As part of our future work, we plan to validate our model on independent datasets, which will strengthen its generalizability and applicability beyond the ADNI dataset.

Conclusion

In conclusion, our study found that standard neuropsychological tests combined with other indicators that differed across phases of disease progression could accurately predict cognitive conversion in patients with AD or at risk thereof. This prognostic information may be

utilized for early upgrade of high-risk patients to more aggressive treatment regimens.

Abbreviations

AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ChEIs	Cholinesterase inhibitors
CSF	Cerebrospinal fluid
PET	Positron emission tomography
Aβ	βamyloid
p-tau	Phosphorylated tau
NfL	Neurofilament light
MMSE	mini mental state examination
ADAS-cog	Alzheimer's disease assessment scale-cognition
AVLT	Auditory verbal learning test
CN	Cognitive normal state
SCD	Subjective cognitive decline
MCI	Mild cognitive impairment
APOE	Apolipoprotein E
PCR	Polymerase chain reaction
HIS	Hachinski ischemic scale
MOCA	Montreal cognitive assessment
AVLT-IM	Auditory verbal learning test-immediate recall
AVLT-L	Auditory verbal learning test-learning
AVLT-IF	Auditory verbal learning test-forgetting
AVLT-PC	Auditory verbal learning test- percent forgetting
CDR	Clinical dementia rating
NPI	Neuropsychiatric inventory
GDS	Geriatric depression scale
FAQ	Functional assessment questionary
LDEL	Logical memory-delayed recall
TRA-B	Trail Making Test B
Simoa	Single molecule array
MRI	Magnetic resonance imaging
CoI	Cognition improved
CNI	Cognition not improved
KNN	K-nearest neighbors
AUC	Area under curve value
TP	True positive
TN	True negative
FN	False negative
FP	False positive
AUPRC	Area under the precision-recall curve
FPR	False positive rates

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13195-025-01686-x .

Supplementary Material 1

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Author contributions

The first draft of the manuscript was written by Siyu Yang who was also responsible for data curation and analysis. Xintong Zhang were responsible for data curation and visualization. Xinyu Du, Peng Yan were responsible for visualization. Jing Zhang, Wei Wang, Jing Wang and Lei Zhang were responsible for methodology. Huaiqing Sun, Yin Liu, Xinran Xu, Yaxuan Di, Jin Zhong and Caiyun Wu were responsible for data processing. Ting Wu, Yu Zheng, and Jan D. Reinhardt contributed to design of the study and provided critical revisions of the manuscript. Ting Wu contributed to funding acquisition, in addition, and was responsible for scientific supervision. All authors contributed intellectually important content and approved the final manuscript.

Funding

This study was funded by the National Natural Science Foundation of China (Grant number of 81772454 and 81971237), the Key Project of Jiangsu Province's Key Research and Development Program (BE2023023-2).

Data availability

The data used in this study are from the ADNI database (http://adni.loni.usc.edu), which is accessible to interested scientists with the ADNI Data Use Agreement.

Declarations

Ethics approval and consent to participate

The study was approved by the institutional review boards of all participating institutions involved: Oregon Health and Science University; University of Southern California; University of California, San Diego; University of Michigan; Mayo Clinic, Rochester, MN, USA; Baylor College of Medicine; Columbia University; Washington University in St. Louis; University of Alabama-Birmingham; Mount Sinai School of Medicine; Rush University Medical Center; Wien Center; The Johns Hopkins University; University of South Florida Health Byrd Alzheimer's Institute; New York University; Duke University Medical Center; University of Pennsylvania; University of Kentucky; University of Pittsburgh; University of Rochester Medical Center; University of California, Irvine; University of Texas Southwestern Medical Center; Emory University; University of Kansas; University of California, Los Angeles; Mayo Clinic, Jacksonville, FL, USA; Indiana University; Yale University School of Medicine; Jewish General Hospital/McGill University; Sunnybrook Health Sciences Centre; University of British Columbia; St. Joseph's Hospital, Ontario, Canada; Northwestern University; Nathan S. Kline Institute for Psychiatric Research; Premiere Research Institute; University of California, San Francisco; Georgetown University; Brigham and Women's Hospital; Stanford University; Banner Sun Health Research Institute; Boston University School of Medicine; Howard University; Case Western Reserve University; University of California, Davis; DENT Neurologic Institute; Parkwood Hospital; University of Wisconsin; University of California, Irvine Brain Imaging Center; Banner Alzheimer's Institute; The Ohio State University; Albany Medical College; University of Iowa; Dartmouth-Hitchcock Medical Center; Wake Forest University Health Sciences Center; Rhode Island Hospital; Cornell Medical Center; Cleveland Clinic Lou Ruvo Center for Brain Health (CCLRBC); Roper St. Francis Hospital; and Butler Hospital Memory and Aging Program. The information on ethical approval and the centres involved in the ADNI study as listed above was obtained from the ADNI Data and Publications Committee. Written informed consent was obtained from all participants or their authorized representatives.

Consent for publication

No applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Neurology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu 210029, China
[2]Department of Neurology, Nanjing Second Hospital, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210003, China
[3]Department of Rehabilitation Medicine, The First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China
[4]China-Australia Joint Research Centre for Infectious Diseases, School of Public Health, Xi'an Jiaotong University Health Science Centre, Xi'an, Shanxi 710061, China
[5]Artificial Intelligence and Modelling in Epidemiology Program, Melbourne Sexual Health Centre, Alfred Health, Melbourne, Australia
[6]Central Clinical School, Faculty of Medicine, Monash University, Melbourne, Australia
[7]Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University, Zhengzhou, Henan 450001, China

[8]Department of Neurology, The Affiliated Brain Hospital of Nanjing Medical University, Nanjing, Jiangsu 210029, China
[9]Institute for Disaster Management and Reconstruction, Sichuan University, Chengdu, Sichuan 610207, China
[10]Swiss Paraplegic Research, Nottwil, Switzerland
[11]Department of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland

Received: 17 September 2024 / Accepted: 28 January 2025
Published online: 13 February 2025

References

1. Jack CR Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. Alzheimer's Dement J Alzheimer's Assoc. 2018;14(4):535–62.
2. Soldan A, Pettigrew C, Fagan AM, Schindler SE, Moghekar A, Fowler C, et al. ATN profiles among cognitively normal individuals and longitudinal cognitive outcomes. Neurology. 2019;92(14):e1567–79.
3. Gao F, Lv X, Dai L, Wang Q, Wang P, Cheng Z et al. A combination model of AD biomarkers revealed by machine learning precisely predicts Alzheimer's dementia: China Aging and Neurodegenerative Initiative (CANDI) study. Alzheimers Dement. 2022.
4. Elmaleh DR, Farlow MR, Conti PS, Tompkins RG, Kundakovic L, Tanzi RE. Developing effective Alzheimer's Disease therapies: clinical experience and future directions. J Alzheimers Dis. 2019;71(3):715–32.
5. Salehipour A, Bagheri M, Sabahi M, Dolatshahi M, Boche D. Combination therapy in Alzheimer's Disease: is it time? J Alzheimers Dis. 2022;87(4):1433–49.
6. Grossberg GT, Tong G, Burke AD, Tariot PN. Present algorithms and Future treatments for Alzheimer's Disease. J Alzheimers Dis. 2019;67(4):1157–71.
7. Cummings JL, Tong G, Ballard C. Treatment combinations for Alzheimer's Disease: current and future Pharmacotherapy options. J Alzheimers Dis. 2019;67(3):779–94.
8. Yu F, Vock DM, Zhang L, Salisbury D, Nelson NW, Chow LS, et al. Cognitive effects of Aerobic Exercise in Alzheimer's Disease: a pilot randomized controlled trial. J Alzheimers Dis. 2021;80(1):233–44.
9. James C, Ranson JM, Everson R, Llewellyn DJ. Performance of Machine Learning algorithms for Predicting Progression to Dementia in Memory Clinic patients. JAMA Netw Open. 2021;4(12):e2136553.
10. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143(6):1920–33.
11. Grueso S, Viejo-Sobera R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alzheimers Res Ther. 2021;13(1):162.
12. Pereira JB, Janelidze S, Stomrud E, Palmqvist S, van Westen D, Dage JL, et al. Plasma markers predict changes in amyloid, tau, atrophy and cognition in non-demented subjects. Brain. 2021;144(9):2826–36.
13. Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Doré V, et al. High performance plasma amyloid-β biomarkers for Alzheimer's disease. Nature. 2018;554(7691):249–54.
14. Brickman AM, Manly JJ, Honig LS, Sanchez D, Reyes-Dumeyer D, Lantigua RA, et al. Plasma p-tau181, p-tau217, and other blood-based Alzheimer's disease biomarkers in a multi-ethnic, community study. Alzheimers Dement. 2021;17(8):1353–64.
15. Pérez-Grijalba V, Romero J, Pesini P, Sarasa L, Monleón I, San-José I, et al. Plasma Aβ42/40 ratio detects early stages of Alzheimer's Disease and correlates with CSF and neuroimaging biomarkers in the AB255 study. J Prev Alzheimers Dis. 2019;6(1):34–41.
16. Chatterjee P, Pedrini S, Doecke JD, Thota R, Villemagne VL, Doré V, et al. Plasma Aβ42/40 ratio, p-tau181, GFAP, and NfL across the Alzheimer's disease continuum: a cross-sectional and longitudinal study in the AIBL cohort. Alzheimers Dement. 2023;19(4):1117–34.
17. Chatterjee P, Pedrini S, Ashton NJ, Tegg M, Goozee K, Singh AK, et al. Diagnostic and prognostic plasma biomarkers for preclinical Alzheimer's disease. Alzheimers Dement. 2022;18(6):1141–54.
18. Jia X, Wang Z, Huang F, Su C, Du W, Jiang H, et al. A comparison of the Mini-Mental State Examination (MMSE) with the Montreal Cognitive Assessment (MoCA) for mild cognitive impairment screening in Chinese middle-aged and older population: a cross-sectional study. BMC Psychiatry. 2021;21(1):485.
19. Siqueira GSA, Hagemann PMS, Coelho DS, Santos FHD, Bertolucci PHF. Can MoCA and MMSE be interchangeable cognitive screening tools? A systematic review. Gerontologist. 2019;59(6):e743–63.
20. Fernaeus SE, Ostberg P, Wahlund LO, Hellström A. Memory factors in Rey AVLT: implications for early staging of cognitive decline. Scand J Psychol. 2014;55(6):546–53.
21. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement. 2012;8(1 Suppl):S1–68.
22. Moroney JT, Bagiella E, Desmond DW, Hachinski VC, Mölsä PK, Gustafson L, et al. Meta-analysis of the Hachinski Ischemic score in pathologically verified dementias. Neurology. 1997;49(4):1096–105.
23. Malloy PF, Cummings JL, Coffey CE, Duffy J, Fink M, Lauterbach EC, et al. Cognitive screening instruments in neuropsychiatry: a report of the Committee on Research of the American Neuropsychiatric Association. J Neuropsychiatry Clin Neurosci. 1997;9(2):189–97.
24. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53(4):695–9.
25. Savage RM, Gouvier WD. Rey Auditory-Verbal Learning Test: the effects of age and gender, and norms for delayed recall and story recognition trials. Archives Clin Neuropsychology: Official J Natl Acad Neuropsychologists. 1992;7(5):407–14.
26. Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry. 1982;140:566–72.
27. Cummings JL. The neuropsychiatric inventory: assessing psychopathology in dementia patients. Neurology. 1997;48(5 Suppl 6):S10–6.
28. Smarr KL, Keefer AL. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for epidemiologic studies Depression Scale (CES-D), geriatric Depression Scale (GDS), hospital anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). Arthritis Care Res. 2011;63(Suppl 11):S454–66.
29. Ito K, Hutmacher MM, Corrigan BW. Modeling of Functional Assessment Questionnaire (FAQ) as continuous bounded data from the ADNI database. J Pharmacokinet Pharmacodyn. 2012;39(6):601–18.
30. Gass CS, Patten B, Penate A, Rhodes A. An enhanced delayed recognition measure for the logical memory subtest of the Wechsler Memory Scale - IV. Appl Neuropsychol Adult. 2022;29(2):279–83.
31. Llinàs-Reglà J, Vilalta-Franch J, López-Pousa S, Calvó-Perxas L, Torrents Rodas D, Garre-Olmo J. Trail Mak Test Assess. 2017;24(2):183–96.
32. Rissin DM, Kan CW, Campbell TG, Howes SC, Fournier DR, Song L, et al. Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. Nat Biotechnol. 2010;28(6):595–9.
33. Qaseem A, Snow V, Cross JT Jr., Forciea MA, Hopkins R Jr., Shekelle P, et al. Current pharmacologic treatment of dementia: a clinical practice guideline from the American College of Physicians and the American Academy of Family Physicians. Ann Intern Med. 2008;148(5):370–8.
34. Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M et al. Autogluon-tabular: robust and accurate automl for structured data. arXiv Preprint arXiv:200306505. 2020.
35. Goldstein H, Healy MJR. The graphical presentation of a Collection of means. J Royal Stat Society: Ser (Statistics Society). 1995;158(1):175–7.
36. Janelidze S, Mattsson N, Palmqvist S, Smith R, Beach TG, Serrano GE, et al. Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. Nat Med. 2020;26(3):379–86.
37. Ocasio E, Duong TQ. Deep learning prediction of mild cognitive impairment conversion to Alzheimer's disease at 3 years after diagnosis using longitudinal and whole-brain 3D MRI. PeerJ Comput Sci. 2021;7:e560.
38. Shen XN, Li JQ, Wang HF, Li HQ, Huang YY, Yang YX, et al. Plasma amyloid, tau, and neurodegeneration biomarker profiles predict Alzheimer's disease pathology and clinical progression in older adults without dementia. Alzheimers Dement (Amst). 2020;12(1):e12104.
39. Shadbahr T, Roberts M, Stanczuk J, Gilbey J, Teare P, Dittmer S, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. Commun Med (Lond). 2023;3(1):139.

Publisher's note