Detecting prodromal Alzheimer's disease with MRI through deep learning

Xinyang Feng[1], Frank A. Provenzano[2,3], Scott A. Small[2,3], for the Alzheimer's Disease Neuroimaging Initiative[†]

1. Department of Biomedical Engineering, Columbia University, New York, NY, 10027, United States

2. Department of Neurology, Columbia University, New York, NY, 10032, United States

3. Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, 10032, United States

[†] 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: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

ABSTRACT

Deep learning applied to MRI for Alzheimer's classification is hypothesized to improve if the deep learning model implicates disease's pathophysiology. The challenge in testing this hypothesis is that large-scale data are required to train this type of model. Here, we overcome this challenge by using a novel data augmentation strategy and show that our MRI-based deep learning model classifies Alzheimer's dementia with high accuracy. Moreover, a class activation map was found dominated by signal from the hippocampal formation, a site where Alzheimer's pathophysiology begins. Next, we tested the model's performance in prodromal Alzheimer's when patients present with mild cognitive impairment (MCI). We retroactively dichotomized a large cohort of MCI patients who were followed for up to 10 years into those with and without prodromal Alzheimer's at baseline and used the dementia-derived model to generate individual 'deep learning MRI' scores. We compared the two groups on these scores, and on other biomarkers of amyloid pathology, tau pathology, and neurodegeneration. The deep learning MRI scores outperformed nearly all other biomarkers, including—unexpectedly—biomarkers of amyloid or tau pathology, in classifying prodromal disease and in predicting clinical progression. Providing a mechanistic explanation, the deep learning MRI scores were found to be linked to regional tau pathology, through investigations using cross-sectional, longitudinal, premortem and postmortem data. Our findings validate that a disease's known pathophysiology can improve the design and performance of deep learning models. Moreover, by showing that deep learning can extract useful biomarker information from conventional MRIs, the advantages of this model extend practically, potentially reducing patient burden, risk, and cost.

INTRODUCTION

Biomarkers can aid in the clinical evaluation of Alzheimer's disease (AD), and biomarkers currently exist for AD's three core neuropathologies—amyloid pathology, tau pathology, and neurodegeneration[1,2]. The first two can be estimated from CSF levels of Aβ and tau, or by direct visualization using PET-sensitive radioligands. Neurodegeneration, a term currently used to encompass neuronal or synaptic loss[3], can be estimated from PET-based measures of parietal cortex metabolism, or MRI-based measurements that reflect the structural integrity of the hippocampal formation.

Deep learning is a subset of machine learning that, in principle, holds promise for MRI-based classification of neurogenerative diseases, including AD[4,5]. Furthermore, while some studies have examined classifying MCI conversion using machine learning frameworks, they have done so using other architectures like SVM[6], examining only up to 36 months[6-8], using clinical information in the model[7-9], and few have examined performance independently against existing biomarkers. We hypothesized that designing a deep learning model that considers AD's known pathophysiology and anatomy would improve the model's classification ability. Because 'cell sickness' occurs first and foremost in the pathophysiology of AD[3,10,11], before dramatic neuronal loss, a classifier is predicted to be improved if it is based on alterations in voxel signal intensity rather than on volume shrinkage. Additionally, informed by the brain's anatomical complexity, particularly the areas whose AD's pathophysiology targets, a classifier is expected to improve if it is based on 3D than on 2D MRI information.

The challenge with a 3D classifier that depends on voxel signal intensity is that its training is estimated to require an unusually large number of scans from cases and controls, more than is typically available for AD. Having access to large-scale datasets is a common challenge for deep learning in all fields, and strategies have been developed for data augmentation[12]. We deploy a novel data augmentation strategy that is particularly well suited for MRI-only datasets, by including scans acquired from the same patient across multiple visits. By training, validating, and testing the classifier at the level of individual subjects, instead of individual scans, we minimize the potential limitations of this approach, namely data leakage.

We elected not to augment data by traditional methods of image perturbation, like rotating or applying transformations, since structural MRI data have well known preprocessing pipelines to spatially align images. We did not include available clinical information, as studies have done prior[7], to avoid a model dependent on information that might be sparse or unavailable, as might be the case of clinical evaluation outside of a carefully controlled and harmonized setting, like ADNI.

In the first series of studies, we used this data augmentation strategy to accumulate a large-scale dataset of MRI scans generated from patients with AD dementia and controls, and from which we could test our hypothesis about a deep learning model that uses an intensity-dependent 3D classifier. Confirming our hypothesis, the model, which generates individual ‘deep learning MRI’ scores reflecting AD probability, was found to classify AD dementia with very high accuracy. Moreover, although voxels from the whole brain were included in the model, the most predictive areas turned out to encompass the hippocampal formation. This anatomical profile supports the biological premise of our classification, potentially placing our deep learning MRI scores within the ‘neurodegeneration’ biomarker category.

While these results were encouraging, AD progresses through a prodromal stage before causing dementia, presenting clinically as mild cognitive impairment (MCI)[13]. Only a subset of patients with MCI has prodromal AD, and in contrast to AD dementia, where a clinical evaluation is often sufficient to diagnose the disease, our ability to diagnose prodromal AD when presented with an MCI patient is currently inadequate. With increased awareness and concern over AD, a growing number of MCI patients are presenting to clinicians wanting to know whether they have prodromal AD, and, if so, how quickly they will progress to dementia. Showing that the deep learning algorithm can address the clinical questions that relate to prodromal AD would not only better validate its classification capabilities, but since derived from conventionally-acquired MRI scans, would potentially expand its clinical utility.

Accordingly, in the second series of studies we set out to test how well the deep learning MRI scores, derived from the deep learning model trained on AD dementia, performs in detecting prodromal AD and in predicting time to dementia progression. Additionally, we compared its performance to other biomarkers of amyloid pathology, tau pathology, and neurodegeneration. Based on the premise of deep learning's classification abilities, we hypothesized that deep learning MRI scores would outperform other MRI-based biomarkers of neurodegeneration. At the same time, given the proposed temporal profile of AD's neuropathology[14], we hypothesized that amyloid or tau biomarkers would outperform the deep learning MRI score in classifying prodromal AD. Additionally, we investigated the link of deep learning MRI scores to amyloid and tau pathology, using cross-sectional, longitudinal, premortem and postmortem data, providing mechanistic explanation for the deep learning MRI score.

The diagnostic cutoffs for all AD biomarkers are traditionally derived from patients in the dementia stage, and biomarkers shift over the disease's progressive course, particularly dynamic during its early stages. Since cutoffs for prodromal AD have not yet been established for any of the biomarkers, the best experimental design with which to test these hypotheses is to clinically follow a large group of MCI patients as they progress to dementia, so that the patients can be retroactively dichotomized into those with and without prodromal AD at baseline. Biomarkers can then be tested to determine which best classifies prodromal AD and which best predicts progression. The challenge with this design is that, based on current estimates, approximately 5 years of clinical follow-up is needed in order to allot sufficient time for the majority of prodromal AD patients to clinically manifest as dementia[15,16]. Here, we were able to implement this experimental design thanks to the Alzheimer's Disease Neuroimaging Initiative (ADNI), which has been acquiring biomarker data in a large population of MCI patients since 2005 and to test the two hypotheses about which biomarker best classifies prodromal AD, and which predicts progression to dementia.

RESULTS

Classifying the dementia stage of Alzheimer's disease

The deep learning model was trained, validated, and tested on 975 MRI scans repeatedly acquired in patients in the dementia stage of AD, versus 1943 MRI scans repeatedly acquired from healthy controls. In the test set, a 'deep learning MRI' score was derived for each scan from the model, with the score reflecting the probability of each scan having AD. A receiver operating characteristic (ROC) analysis revealed that the deep learning MRI scores accurately classified AD dementia vs. healthy controls with an AUROC (area under the receiver operating characteristics curve) of 0.973 (Fig. 1a).

Next, we generated an AD 'class activation map' to determine whether the deep learning MRI scores derived from the model were regionally dominated. We find that the deep learning MRI scores are dominated by alterations in voxel signal intensity that localized to anterior medial temporal lobe, in the vicinity of the anterior entorhinal cortex and hippocampus (Fig. 1b). We note that while the class activation map localized to the left more than the right anterior medial temporal lobe, in agreement with previous findings[17-19], contralateral areas emerged with lowered thresholding.

Classifying the prodromal stage of Alzheimer's disease

From ADNI, we identified a cohort of participants who were diagnosed with MCI at baseline and who had a complete set of CSF amyloid and tau biomarkers and structural MRI (N = 582; the inclusionary and exclusionary algorithm is illustrated in Fig. S1).

Among these, 205 participants progressed to AD dementia at follow up ('MCI progression' group), and thus had prodromal AD at baseline, while 179 participants remained MCI stable for at least 4 years ('MCI stable' group) (Fig. 2). The dementia-derived deep learning classifier was used to generate deep learning MRI scores on each individual case.

ROC analyses revealed that the deep learning MRI score outperformed all other biomarkers in classifying the MCI-stable from the MCI-progression group (Fig. 3). The AUROC of deep learning MRI score was 0.788 (Accuracy at Youden (ACC)=75%), superior to CSF Aβ (AUROC=0.702 ACC=66.7%, significantly lower than the deep learning MRI score, p=0.0141), CSF tau (AUROC=0.682, ACC=66.4%, p=0.0161), CSF tau/Aβ (AUROC=0.703, ACC=68.5%, p=0.0161); superior to MRI-based measures of hippocampal volume (AUROC=0.733, ACC=67.7%, p=0.0484), entorhinal cortex volume (AUROC=0.64, ACC=62.5%, p=2.01E-6), and entorhinal cortex thickness (AUROC=0.685, ACC=64.1%, p=1.71E-4); and, finally, superior to Mini-Mental State Exam (AUROC=0.648, ACC=63.3%, p=6.70E-5), and to neuropsychological measure most sensitive to the early stages of AD, the RAVLT retention score[20] (AUROC=0.686, ACC=67.7%, p=2.28E-3).

Additionally, the deep learning MRI score was found to outperform or perform as well when tested in a subset of participants in whom additional PET-based biomarkers were available -- FDG-PET that by measuring parietal cortex metabolism is considered a biomarker of neurodegeneration[21], and AV45-PET, which by using an amyloid radioligand is a biomarker of amyloid pathology[22]. In this subset, the deep learning MRI score classified prodromal AD with an AUROC=0.815 (ACC=78.6%), compared to the

AUROC of 0.782 (for PDG-PET (ACC=75.4%) and 0.751 (ACC=71.4%) for amyloid-PET, although the differences were not statistically significant (Fig. 3, bottom panel).

Predicting progression to Alzheimer's disease dementia

Survival analyses were performed to determine which biomarker best predicted progression to AD dementia among the MCI groups. Results revealed that compared to other biomarkers, the deep learning MRI score best predicted time to conversion to AD dementia, as illustrated by the survival curves of high and low deep learning MRI scores and tau/Aβ ratios (Fig. 4). The deep learning MRI scores showed better prediction capability (|z|=11.0, p=4.35E-28) than CSF biomarkers of amyloid and tau pathology (Aβ |z|=6.37, p=1.87E-10, tau |z|=5.70, p=1.18E-08, tau/Aβ |z|=5.41, p=6.29E-08); than MRI-based biomarkers of neurodegeneration (hippocampal volume |z|=8.80, p=1.35E-18, entorhinal volume |z|=6.02, p=1.75E-09, entorhinal thickness |z|=7.42, p=1.21E-13); and, than behavioral measures (MMSE |z|=5.72, p=1.07E-08, RAVLT retention |z|=6.88, p=6.12E-12). Similarly, in the subset in whom the additional PET biomarkers were available the deep learning MRI score (|z|=9.04, p=1.40E-19) outperformed or performed as well as FDG-PET (|z|=9.11, p=8.14E-20) and AV45-PET |z|=7.12, p=1.04E-12).

Correlations with amyloid pathology and tau pathology

Correlational analyses were performed to determine whether the deep learning MRI score was correlated more with amyloid pathology or tau pathology. Cross-sectionally, we found that while the deep learning MRI score showed a stronger correlation with CSF tau (r=0.225, p=9.00E-6), it also correlated with CSF Aβ (r=-0.190, p=1.86E-4). Longitudinally, however, changes in the deep learning MRI scores over time were significantly associated with changes in CSF tau (r=-0.205, p=1.50E-3), but not with changes in CSF Aβ (r=-8.18E-3, p=0.900).

Next, in a subsample with available postmortem data, we correlated the deep learning MRI score with neuropathological evidence of amyloid pathology, as indicated by the Thal staging[23], or tau pathology indicated by Braak staging[24]. The deep learning MRI scores were found to associate more with tau pathology (with an MRI-autopsy interval below 2 years, Braak staging: r=0.397, p=7.70e-3; Thal staging: r=0.196, p=0.203) (Fig. 5 bottom panel). To further explore the regionality of this relationship, we found that the deep learning MRI score correlated with tau levels mapped by tau-PET, with strong correlations observed with tau pathology in the entorhinal cortex (r=0.449, p=1.66E-15).

DISCUSSION

The level of performance achieved by our deep learning model in classifying AD dementia supports our hypothesis that a disease's pathophysiology should be considered when evaluating performance as well as justifying the data augmentation strategy used. Further validating the assumptions, design, and implementation of our model is the fact that, despite incorporating information from the whole brain, the class activation map was dominated by signal in the anterior entorhinal cortex and hippocampus, precisely where AD pathophysiology begins[3,17-19,24].

Stronger validation of the deep learning model was provided by the second series of studies when the dementia-derived classifier was applied to the prodromal stages of AD. Supporting the first hypothesis of this study, we found that our deep learning MRI scores outperformed other MRI-based measures of neurodegeneration in both classifying prodromal AD and in predicting progression to dementia. Refuting the second hypothesis, we found that the deep learning MRI scores performed at least as well and typically outperformed biomarkers of amyloid and tau pathology.

We do not consider this unexpected finding a challenge to the primacy of amyloid and tau pathology in the neuropathological progression of AD[25]. The deep learning MRI scores were found strongly linked to tau pathology in the entorhinal cortex, a region where AD pathology begins[24], and its superior performance likely reflects this sensitivity. It is possible, therefore, that tau-PET would outperform the deep learning MRI score and other biomarkers. ADNI, however, has only begun acquiring tau-PET in 2015, and there is currently insufficient data to test this prediction in our experimental design.

Future analyses from ADNI and other long-term PET studies will be able to test this prediction.

The observation that the deep learning MRI scores outperformed biomarkers of amyloid and tau pathology in predicting time to dementia is less surprising. As a biomarker of neurodegeneration, this finding agrees with prior studies[26] and with the current model for the temporal sequence of AD's neuropathology[25]. Since in this scheme neurodegeneration occurs last, accurate biomarkers of it are more proximal to the development of dementia.

The strength of our prodromal AD study is that by relying on progression to AD dementia as a way to retroactively identify patients with prodromal AD, we overcame the limitation that precise biomarker cutoffs for prodromal AD have not yet been established. We designed the analysis based on prior studies that suggest that the majority of MCI patients with prodromal AD will progress to dementia within 4-5 years[15], an assumption confirmed in our study. Furthermore, approximately half of the MCI cohort ended up having prodromal AD, which agrees with previous approximations[27]. Still, a potential weakness of our study is the possibility that a minority of patients in the stable MCI category are harboring prodromal AD at baseline. The number of misclassified patients is likely to be low[27], and so this potential imprecision would not be expected to significantly alter our results. Tracking stable MCI patients for longer periods might address this concern, but would in fact raise a new one: when tracking patients for a decade or more, particularly given the high incidence of AD in older populations, some are expected to develop AD *de novo* after the baseline evaluation. We can conclude that our findings and their conclusions are beyond reproach for a 5-

year time window after initial evaluation, a clinically meaningful epoch for both patients and clinicians.

Our study provides the proof-of-principle that imaging-based deep learning models that are examined in concert with a disease's pathophysiology will yield a highly accurate model and improve performance in prognosticating disease. Showing that deep learning can enhance the utility of MRI in prodromal AD is the more important clinical implication of this study. Ordering "neuroimaging studies"[28] is the current standard of care when evaluating a patient with MCI suspected of having AD, most typically the conventional MRIs from which the deep learning MRI scores were derived. The rationale for this recommendation and its routine clinical implementation is not to 'rule in' AD, but rather to exclude other non-neurodegenerative causes of dementia, such as strokes, bleeds, and tumors. Deep learning algorithms that can extract useful information for the purposes of prodromal AD detection, from conventional MRIs that have in any case been acquired, has the additional advantages of reducing patient burden and cost incurred by lumbar punctures, injection of radioactive ligands, or other additional testing.

METHODS

Participants in the Alzheimer's disease dementia study

All data were obtained from ADNI, a multi-site observational study, which were acquired in accordance with each site's respective Institutional Review Board, including obtaining written consent acquired from each participant. We included 2918 scans ($N_{\text{healthy control}}$ = 1943, N_{AD} = 975) from 626 subjects as training set, 382 scans ($N_{\text{healthy control}}$ = 251, N_{AD} = 131) from 80 subjects as validation set, and 325 scans ($N_{\text{healthy control}}$ = 229, N_{AD} = 96) from 80 subjects as test set.

Our data augmentation method of using scans from multiple visits of the same participant requires dealing with two problems: data leakage and disease progression. Data leakage is the problem of including different scans from the same participant in the training and test set, the trained model might make the prediction by matching the subject instead of extracting disease relevant patterns. In this study, the training, validation and test sets were partitioned at subject level to ensure non-overlapping subjects. Disease progression is the problem that the diagnosis status of subjects might change during follow-up visits, and the diagnosis at scan time might be different from the baseline label. In this study, we labeled all the scans with their cross-sectional diagnosis at scan time.

Participants in the 'Mild Cognitive Impairment' study

From ADNI we identified a cohort of participants who were diagnosed with MCI at baseline and who had a complete set of CSF amyloid and tau biomarkers and structural

MRI (N = 582; the inclusionary and exclusionary algorithm is illustrated in Fig. S1). Among these, 205 participants progressed to AD dementia at follow up ('MCI progression' group), and 179 participants remained MCI stable for at least 4 years ('MCI stable' group). The time distribution and demographics of these two groups are shown in Fig. 2.

The deep learning MRI score

The deep learning model used in this study is a three-dimensional convolutional neural network (3D CNN) model with five convolutional stages and one fully connected layer with sigmoid output[5]. Each convolutional stage consists of two convolutional layers with rectified linear unit (ReLU) activation function, a batch normalization operation and a max pooling layer. The model was optimized using ADAM method with cross-entropy loss, using a learning rate of 2e-5 determined through grid search. The model was trained on the brain-extracted T1-weighted structural MRI scans from the ADNI cohort to classify patients in the dementia stage of AD versus healthy control subjects. To evaluate the regional contribution to AD classification, we generated a 3D class activation map, which visualizes the predictive regions in deep learning classification models[29,30].

We applied the model trained to classify AD dementia versus healthy controls to the baseline scans of patients diagnosed with MCI. The continuous output from the model is reflective of the progressive structural patterns of AD pathology. We refer to it as a 'deep learning MRI' score. All analyses were performed using this score.

Amyloid and Tau Biomarkers

CSF biomarkers: CSF tau levels, reflective of neurofibrillary tangle, and CSF Aβ levels, reflective of amyloid pathology, were included in the analysis[31]. Additionally, the tau/Aβ ratio, which has been shown to best capture AD[32], was also included[33]. CSF was acquired at individual ADNI sites in accordance to the ADNI acquisition protocols and analyzed as previously described[33]. The median values provided by ADNI were used.

***PET measures*:** In a subset of participants ($N_{\text{MCI-progression}}$ = 94, $N_{\text{MCI-stable}}$ = 154), amyloid pathology was also estimated with PET, mapping amyloid burden with the amyloid-binding radioligand AV45. The composite AV45-PET score provided by ADNI[34] was used in the analyses, which is based on the average AV45 SUVR (standard uptake value ratio) of the frontal, anterior cingulate, precuneus, and parietal cortex relative to the cerebellum[35].

Neurodegeneration Biomarkers

MRI morphometry: FreeSurfer 6.0[36,37] was used to segment the structural MRI scans and derive regional morphometric measures. Hippocampal (HC) volume, entorhinal cortex (EC) volume, and entorhinal cortex thickness were used as structural integrity measures of the hippocampal formation. Hippocampal and entorhinal cortex volume were normalized by the intra-cranial volume (ICV).

***PET measures*:** In a subset of participants ($N_{\text{MCI-progression}}$ = 94, $N_{\text{MCI-stable}}$ = 154), neurodegeneration was also estimated with PET using fluorodeoxyglucose (FDG). The composite FDG score provided by ADNI[34] was used in the analyses, which is based on the average FDG uptake of angular, temporal, and posterior cingulate[21].

Additional Measures

Behavioral and neuropsychological measures: The MMSE (Mini-Mental State Examination) score and RAVLT (Rey Auditory Verbal Learning Test) retention scores were used in the analysis. The RAVLT retention score measures the number of delayed recalled words divided by the number of words learned in the last learning trial (trial 5) and has been found to be one of the most sensitive to AD[20].

Neuropathology**:** Among subjects with postmortem neuropathology data, 44 cases were identified who had an MRI within two years prior to death, and 29 cases were identified who had MRI within one year prior to death. DLMRI scores were derived from the last antemortem MRI scans in this cohort. An association was investigated between DLMRI score and the neuropathologically-derived Braak stage, which reflects neurofibrillary tangles[24], and the Thal phase, which reflects amyloid plaques[23].

Tau-PET**:** ADNI began acquiring PET scan using the AV1451 radioligand, which binds neurofibrillary tangles[38], in the late phase of ADNI2 and resumed in ADNI3. Due to the smaller number of subjects with available longitudinal tau-PET data or follow-up visits, cross-sectional analyses on these subjects (N = 296) using the regional AV1451 retention levels provided by ADNI[34] were performed.

Statistical analysis

***ROC analysis*:** A receiver operating characteristic (ROC) analysis was used to determine the accuracy of the deep learning MRI score in prodromal AD classification, i.e. MCI-stable and MCI-progression classification, using standardized residuals controlling for age, sex, and APOE ε4 frequency with linear regression. The DeLong test[39] was used to test for the significance of the differences in the AUROCs (area under the ROC curve) between DLMRI score and other measures using the pROC R package[40].

***Survival analysis*:** Cox proportional hazards regression models were fit to examine the association between each baseline measure and time to conversion to AD dementia from MCI, controlling for age, sex, and APOE ε4 frequency, using the survival R package[41]. MCI-stable participants are included in the models as censored data with the last visit as the censored point. The high-risk and low-risk survival curves were generated with the 75% percentile and 25% percentile of the observed measures, respectively.

***Longitudinal analysis*:** The longitudinal association between DLMRI score and CSF biomarkers was studied by examining the deviation from baseline measurements for each participant over time. From the ‘MCI-progression’ and ‘MCI-stable’ group, we further identified participants that had at least one follow-up of both MRI and CSF, and collapsed them into a group for longitudinal analysis (n=238). The changes in either CSF biomarker or DLMRI score of all follow-up visits from baseline were used to estimate the slope β of the change in tau (Δtau), Aβ (ΔAβ), and tau/Aβ ratio (Δtau/Aβ) versus the change in DLMRI score (ΔDLMRI) for each participant using linear regression through the origin. Each participant was represented by the point based on the last follow-up visit's ΔDLMRI$_{last}$ (x-coordinate) and the fitted change $\beta\Delta$DLMRI$_{last}$ (y-coordinate) of the respective measure. The last follow-up visit was used to anchor the representation of the participant in order to reflect the full follow-up. A correlation analysis was performed across participants. A linear regression model was fit across participants and illustrated.

***Correlational analysis*: A partial correlation was performed between baseline DLMRI score and CSF biomarkers, regional tau-PET measures, controlling for age, sex, and APOE ε4 frequency. As the Braak stage of neurofibrillary tangles and the Thal phase of amyloid plagues are both rank ordinal measures, we correlated the DLMRI score with the neuropathological measures using Spearman correlation.

ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DISCLOSURES

FAP is a consultant for and equity holder of Imij Technologies. SAS serves on the scientific advisory board of Meira GTX and is an equity holder in Imij Technologies. XF, FAP and SAS have applied for a provisional patent on neuroimaging-based diagnosis.

AUTHORS’ CONTRIBUTIONS

XF, FAP, SAS contributed to literature search, figures, study design, data interpretation, and writing. XF contributed to data collection and data analysis.

REFERENCES

1. Jack, Clifford R. & Holtzman, David M. Biomarker Modeling of Alzheimer's Disease. *Neuron* **80**, 1347-1358 (2013).
2. Olsson, B.*, et al.* CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology* **15**, 673-684 (2016).
3. Khan, U.A.*, et al.* Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease. *Nature Neuroscience* **17**, 304-311 (2014).
4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
5. Feng, X., Yang, J., Lipton, Z.C., Small, S.A. & Provenzano, F.A. Deep Learning on MRI Affirms the Prominence of the Hippocampal Formation in Alzheimer's Disease Classification. *bioRxiv*, 456277 (2018).
6. Moradi, E.*, et al.* Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* **104**, 398-412 (2015).
7. Spasov, S.*, et al.* A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage* **189**, 276-287 (2019).
8. Basaia, S.*, et al.* Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage. Clinical* **21**, 101645 (2019).
9. Tong, T.*, et al.* A Novel Grading Biomarker for the Prediction of Conversion From Mild Cognitive Impairment to Alzheimer's Disease. *IEEE Trans Biomed Eng* **64**, 155-165 (2017).
10. Small, S.A. Alzheimer disease, in living color. *Nature Neuroscience* **8**, 404-405 (2005).
11. Selkoe, D.J. Alzheimer's Disease Is a Synaptic Failure. *Science (New York, N.Y.)* **298**, 789-791 (2002).
12. Krizhevsky, A., Sutskever, I. & Hinton, G.E. Imagenet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems (NIPS)* 1097-1105 (2012).
13. Petersen, R.C.*, et al.* Mild Cognitive Impairment: Clinical Characterization and Outcome. *Archives of Neurology* **56**, 303-308 (1999).
14. Jack, C.R.*, et al.* A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology* **87**, 539-547 (2016).
15. Visser, P.J., Kester, A., Jolles, J. & Verhey, F. Ten-year risk of dementia in subjects with mild cognitive impairment. *Neurology* **67**, 1201-1207 (2006).
16. Mitchell, A.J. & Shiri-Feshki, M. Rate of progression of mild cognitive impairment to dementia – meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica* **119**, 252-265 (2009).
17. Yushkevich, P.A.*, et al.* Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping* **36**, 258-287 (2015).
18. Maass, A.*, et al.* Entorhinal Tau Pathology, Episodic Memory Decline, and Neurodegeneration in Aging. *The Journal of Neuroscience* **38**, 530-543 (2018).

19. Miller, M.I., *et al.* The diffeomorphometry of temporal lobe structures in preclinical Alzheimer's disease. *NeuroImage: Clinical* **3**, 352-360 (2013).
20. Chang, Y.L., *et al.* Brain substrates of learning and retention in mild cognitive impairment diagnosis and progression to Alzheimer's disease. *Neuropsychologia* **48**, 1237-1247 (2010).
21. Landau, S.M., *et al.* Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiology of Aging* **32**, 1207-1218 (2011).
22. Clark, C.M., *et al.* Use of Florbetapir-PET for Imaging β-Amyloid Pathology. *Jama* **305**, 275-283 (2011).
23. Thal, D.R., Rüb, U., Orantes, M. & Braak, H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD. *Neurology* **58**, 1791-1800 (2002).
24. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica* **82**, 239-259 (1991).
25. Jack, C.R., *et al.* Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology* **12**, 207-216 (2013).
26. Vemuri, P., *et al.* MRI and CSF biomarkers in normal, MCI, and AD subjects: Predicting future clinical change. *Neurology* **73**, 294-301 (2009).
27. Vos, S.J., *et al.* Prevalence and prognosis of Alzheimer's disease at the mild cognitive impairment stage. *Brain : a journal of neurology* **138**, 1327-1338 (2015).
28. Albert, M.S., *et al.* The diagnosis of mild cognitive impairment due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* **7**, 270-279 (2011).
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
30. Selvaraju, R.R., *et al.* Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. in *IEEE International Conference on Computer Vision (ICCV)* 618-626 (2017).
31. Tapiola, T., *et al.* Cerebrospinal fluid {beta}-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. *Arch Neurol* **66**, 382-389 (2009).
32. Fagan, A.M., *et al.* Cerebrospinal fluid tau/beta-amyloid(42) ratio as a prediction of cognitive decline in nondemented older adults. *Arch Neurol* **64**, 343-349 (2007).
33. Shaw, L.M., *et al.* Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology* **65**, 403-413 (2009).
34. Jagust, W.J., *et al.* The Alzheimer's Disease Neuroimaging Initiative 2 PET Core: 2015. *Alzheimer's & dementia : the journal of the Alzheimer's Association* **11**, 757-771 (2015).
35. Landau, S. & Jagust, W. Florbetapir processing methods. (http://adni.loni.usc.edu, 2015).
36. Fischl, B., *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341-355 (2002).
37. Fischl, B., *et al.* Automatically parcellating the human cerebral cortex. *Cerebral Cortex* **14**, 11-22 (2004).

38. Marquié, M., *et al.* Validating novel tau positron emission tomography tracer [F-18]-AV-1451 (T807) on postmortem brain tissue. *Annals of Neurology* **78**, 787-800 (2015).
39. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845 (1988).
40. Robin, X., *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* **12**, 77 (2011).
41. Therneau, T.M. & Grambsch, P.M. *Modeling survival data: extending the Cox model*, (Springer Science & Business Media, 2013).

FIGURE LEGENDS

Figure 1. Classifying Alzheimer's disease in its dementia stage.

The 'receiver operating characteristic' curve shows that the deep learning MRI score applied to the test set of Alzheimer's disease (AD) dementia scans vs. healthy controls scans classified AD dementia with high accuracy (panel 'a'). The class activation map, reflective of the regional contributions to the deep learning MRI scores, localized to the left anterior medial temporal lobe in the vicinity of the entorhinal cortex and hippocampus, where Alzheimer's pathophysiology begins.

Figure 2. Distribution and demographics of subjects in the 'mild cognitive impairment' study

Distribution frequencies of the participants with amnestic mild cognitive impairment (MCI) at baseline, who either remained stable (MCI stable) or progressed to Alzheimer's dementia (MCI progression), organized by latest follow-up years and conversion years. The dark blue bars indicate participants included in the study. Demographic and baseline biomarker data are listed in the table for the MCI stable and MCI progression groups.

Figure 3. Classifying Alzheimer's disease in its prodromal stage

By comparing the 'MCI stable' to the 'MCI progression' groups, ROC curves show that the deep learning MRI (DLMRI) scores were superior in classifying prodromal Alzheimer's disease (indicated in red). The deep learning MRI scores outperformed (left panel) CSF measures of Aβ, tau, or tau/Aβ; MRI measures of hippocampal (HC) or entorhinal cortex (EC) volume or thickness; clinical measures using the modified mental status exam (MMSE) or the retention of the Rey Auditory Verbal Learning Task (RAVLT) (left panel). In a smaller subset, the deep learning MRI scores (right panel) outperformed PET measures of amyloid using the AV45 radioligand or metabolism using fluorodeoxyglucose (FDG). Specific area under the curve (AUROC) values for each measure, and statistical probability values for each comparison, are shown in the table.

Figure 4. Predicting progression to Alzheimer's Dementia

Survival analyses were performed comparing the deep learning MRI scores to other measures, and example curves illustrate that the deep learning MRI score (left panel) outperforming the CSF measure of the tau/Aβ ratio (right panel). The high risk (indicated by red) and low risk (indicated by blue) curves were fitted from 75% and 25% percentile of the measures respectively. The shaded area indicates the 95% confidence interval.

The deep learning MRI scores outperformed CSF Aβ, tau, or tau/Aβ, MRI-derived measures of hippocampal volume, entorhinal cortex volume, and entorhinal thickness; behavioral measures, Mini-Mental State Exam (MMSE), and RAVLT retention; and, when available, PET measures of amyloid using the AV45 radioligand or metabolism using fluorodeoxyglucose (FDG).

Figure 5. The deep learning MRI score correlates with tau pathology

The scatter plots illustrate the relationship between changes over time in the deep learning MRI scores vs. changes in CSF Aβ (left panel), changes in CSF tau (middle panel) and changes in CSF tau/Aβ (right panel). Each data point indicates one participant's change of last deep learning MRI score from baseline ($\Delta DLMRI_{last}$), plotted against their fitted change in biomarker measures at $\Delta DLMRI_{last}$ with the slope estimated from all follow-up visits (see Methods). The black solid lines are the linear fits across participants, showing that changes in the deep learning MRI score is most strongly correlated with changes in tau over time. The table lists the correlations between antemortem deep learning MRI scores to postmortem-derived Braak stage of neurofibrillary tangles and the Thal phase of amyloid plaques, with an MRI-autopsy interval below either 1 year and 2 years, showing that the deep learning MRI scores are most strongly correlated with tau pathology.

Figure S1. Participant selection flow-chart

FIGURES

Figure 1

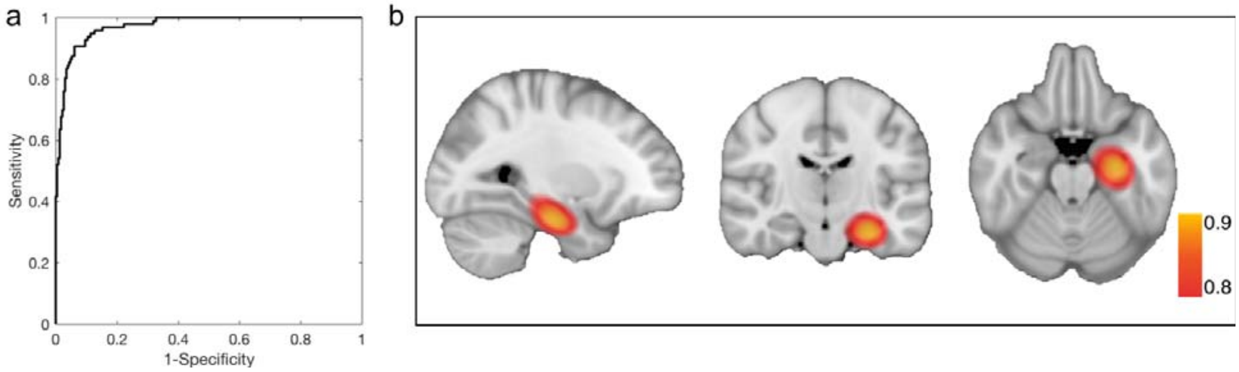
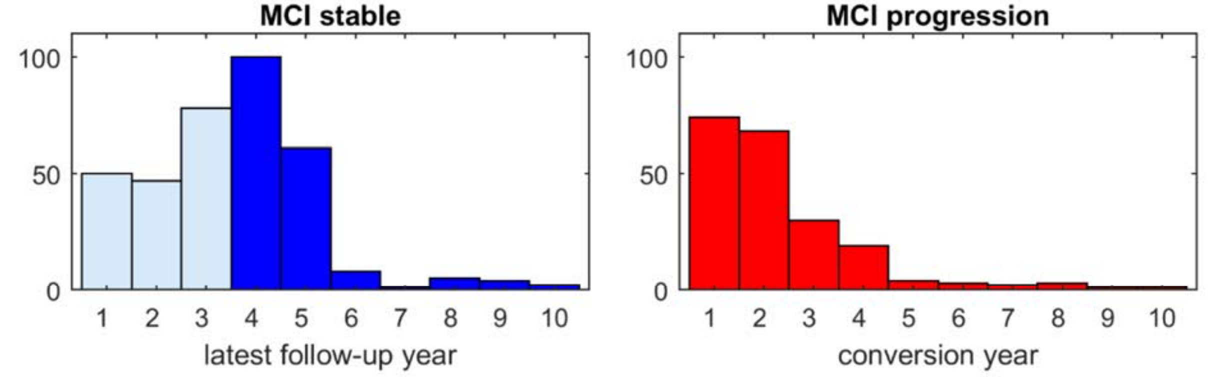


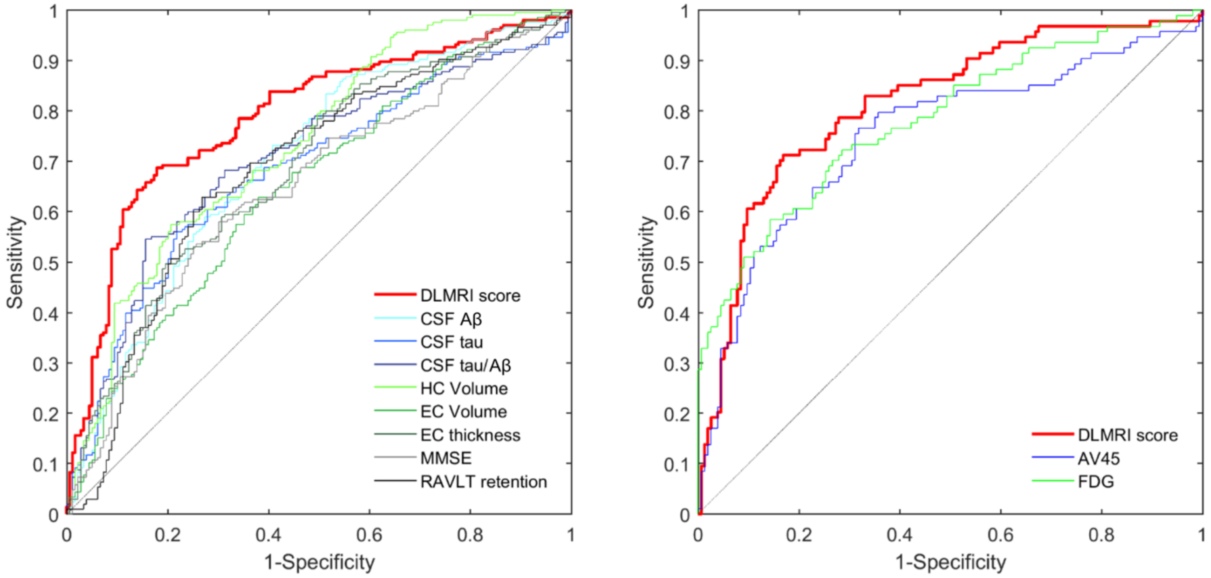
Figure 2



Main dataset	**MCI stable N = 179**	**MCI progression N = 205**	**Total N = 384**
age	71.3 ± 7	73.5 ± 7.1	72.5 ± 7.1
sex M/F (%M)	107/72 (59.8)	124/81 (60.5)	231/153 (60.2)
APOE ε4 frequency (2/1/0)	13/52/114	33/101/71	46/153/185
Last visit FU year	4.74 ± 1.2	-	-
conversion year	-	2.2 ± 1.7	-
DLMRI score	0.385 ± 0.165	0.622 ± 0.196	0.511 ± 0.217
Aβ	190.3 ± 51.3	144 ± 39.8	165.6 ± 51
tau	74.1 ± 43.0	115.3 ± 56.4	96.1 ± 54.6
tau/Aβ	0.457 ± 0.395	0.881 ± 0.532	0.684 ± 0.518
EC thickness	3.34 ± 0.4	3.03 ± 0.43	3.18 ± 0.44
EC volume (/ICV)	2.62E-03 ± 5.15E-04	2.31E-03 ± 4.95E-04	2.46E-03 ± 5.27E-04
HC volume (/ICV)	4.81E-03 ± 6.88E-04	4.14E-03 ± 6.00E-04	4.45E-03 ± 7.25E-04
MMSE	28.1 ± 1.7	26.9 ± 1.8	27.5 ± 1.8
RAVLT retention	49.6 ± 30.8	24.6 ± 29.3	36.2 ± 32.5

PET subset	**MCI stable N = 154**	**MCI progression N = 94**	**Total N = 248**
age	70.9 ± 7	72.8 ± 6.6	71.6 ± 6.9
sex M/F (%M)	88/66 (57.1)	52/42 (55.3)	140/108 (56.5)
APOE ε4 frequency (2/1/0)	11/44/99	17/50/27	28/94/126
Last visit FU year	4.43 ± 0.55	-	-
conversion year	-	1.91 ± 1.16	-
DLMRI score	0.376 ± 0.163	0.635 ± 0.190	0.474 ± 0.214
FDG	1.31 ± 0.11	1.16 ± 0.11	1.26 ± 0.13
AV45	1.15 ± 0.19	1.4 ± 0.21	1.24 ± 0.23

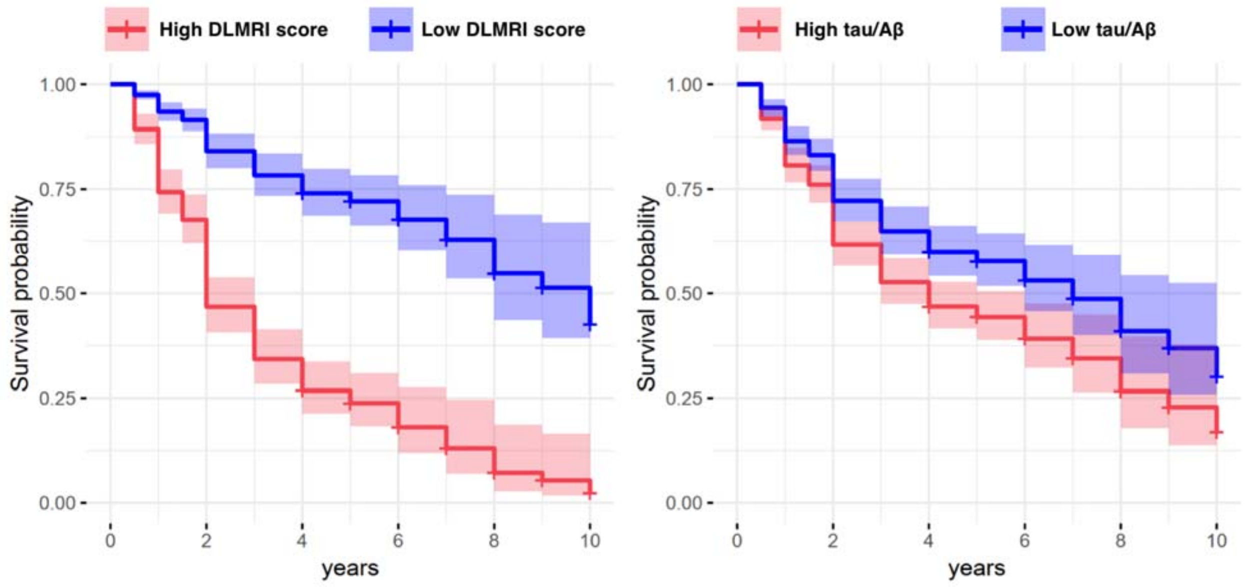
Figure 3

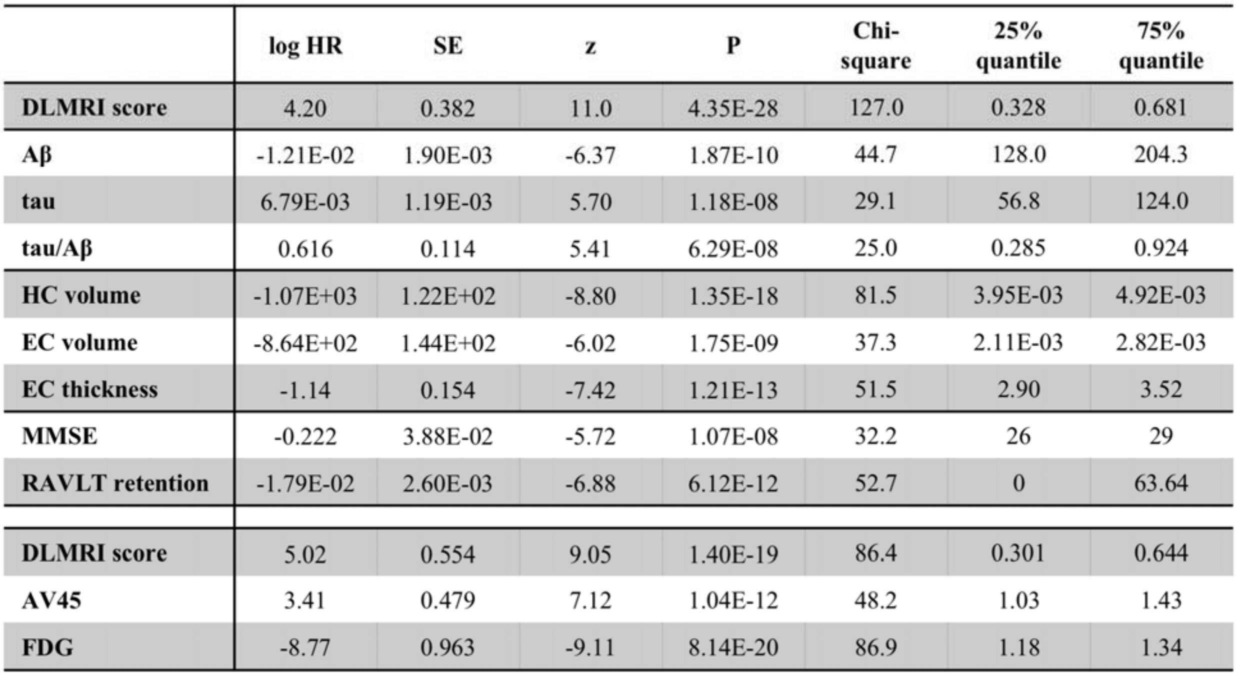


	DLMRI score	Aβ	tau	tau/Aβ	HC volume	EC volume	EC thickness	MMSE	RAVLT retention
AURUC	0.788	0.702	0.682	0.703	0.733	0.648	0.685	0.648	0.686
p-value	-	0.0141	2.47E-3	0.0161	0.0484	2.01E-6	1.71E-4	6.70E-5	2.28E-3

DLMRI score	AV45	FDG
0.815	0.751	0.782
-	0.154	0.330

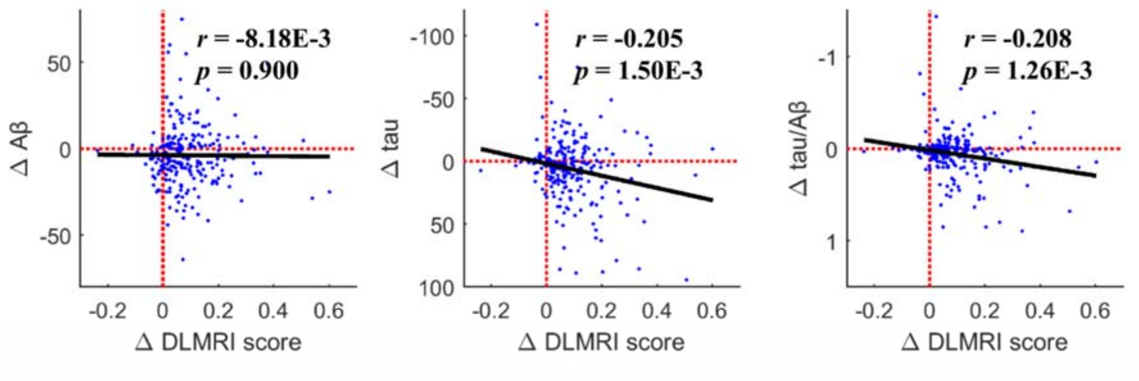
Figure 4





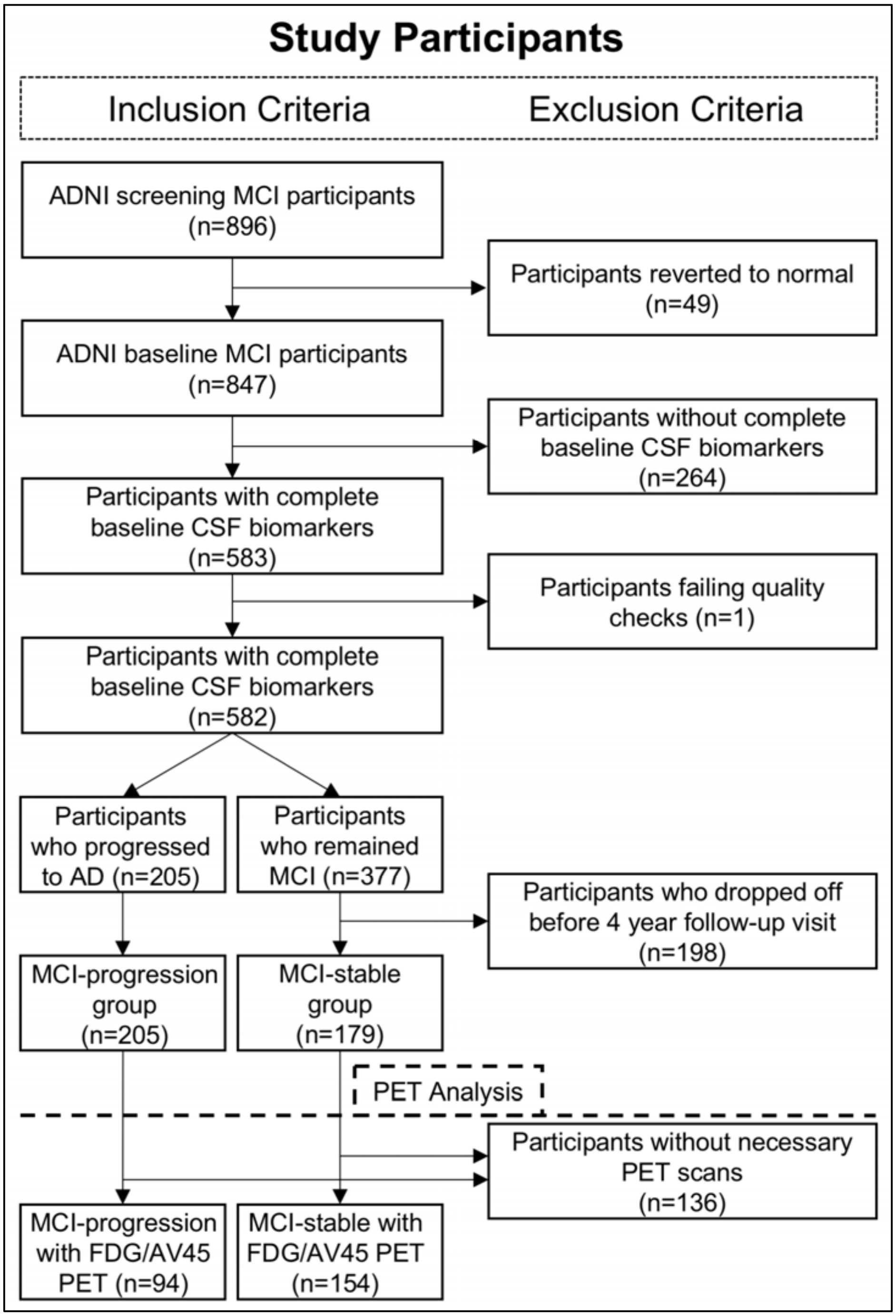
	log HR	SE	z	P	Chi-square	25% quantile	75% quantile
DLMRI score	4.20	0.382	11.0	4.35E-28	127.0	0.328	0.681
Aβ	-1.21E-02	1.90E-03	-6.37	1.87E-10	44.7	128.0	204.3
tau	6.79E-03	1.19E-03	5.70	1.18E-08	29.1	56.8	124.0
tau/Aβ	0.616	0.114	5.41	6.29E-08	25.0	0.285	0.924
HC volume	-1.07E+03	1.22E+02	-8.80	1.35E-18	81.5	3.95E-03	4.92E-03
EC volume	-8.64E+02	1.44E+02	-6.02	1.75E-09	37.3	2.11E-03	2.82E-03
EC thickness	-1.14	0.154	-7.42	1.21E-13	51.5	2.90	3.52
MMSE	-0.222	3.88E-02	-5.72	1.07E-08	32.2	26	29
RAVLT retention	-1.79E-02	2.60E-03	-6.88	6.12E-12	52.7	0	63.64
DLMRI score	5.02	0.554	9.05	1.40E-19	86.4	0.301	0.644
AV45	3.41	0.479	7.12	1.04E-12	48.2	1.03	1.43
FDG	-8.77	0.963	-9.11	8.14E-20	86.9	1.18	1.34

Figure 5



MRI-autopsy interval	1 year (N = 29)		2 years (N = 44)	
	corr-coef	p-value	corr-coef	p-value
Braak stage (neurofibrillary tangles)	0.402	0.0305	0.397	7.70E-3
Thal phase (amyloid plaques)	0.239	0.2119	0.196	0.2030

Figure S1


Study Participants
Inclusion Criteria
Exclusion Criteria
ADNI screening MCI participants (n=896)
Participants reverted to normal (n=49)
ADNI baseline MCI participants (n=847)
Participants without complete baseline CSF biomarkers (n=264)
Participants with complete baseline CSF biomarkers (n=583)
Participants failing quality checks (n=1)
Participants with complete baseline CSF biomarkers (n=582)
Participants who progressed to AD (n=205)
Participants who remained MCI (n=377)
Participants who dropped off before 4 year follow-up visit (n=198)
MCI-progression group (n=205)
MCI-stable group (n=179)
PET Analysis
Participants without necessary PET scans (n=136)
MCI-progression with FDG/AV45 PET (n=94)
MCI-stable with FDG/AV45 PET (n=154)